Phantoms Never Die: Living with Unreliable Mortality Data

Andrew J.G. Cairns[a]

David Blake[b]

Kevin Dowd[c]

Amy Kessler[d]

First version: December 2013
This version: July 1, 2014

Abstract

The analysis of national mortality trends is critically dependent on the quality of the population, exposures and deaths data that underpin death rates. This paper, using England & Wales population data by way of example, develops a framework that allows us to assess data reliability and identify anomalies. First, we propose a set of graphical diagnostics that help to pinpoint anomalies. Second, we develop a simple model that allows us to quantify objectively the size of any anomalies.

An important conclusion is that bigger anomalies can often be linked to uneven patterns of births in cohorts born in the distant past, leading to errors of more than 9% in the estimated size of some England & Wales birth cohorts. We propose methods that can use the births data from these cohorts to improve estimates of the underlying population exposures.

Keywords: Mortality data, deaths, population, exposures, cohort-births-deaths exposures methodology, convexity adjustment ratio, graphical diagnostics.

[a]Corresponding author. Maxwell Institute for Mathematical Sciences, Edinburgh, and Department of Actuarial Mathematics and Statistics, Heriot-Watt University, Edinburgh, EH14 4AS, United Kingdom. WWW: http://www.macs.hw.ac.uk/∼andrewc E-mail: A.J.G.Cairns@hw.ac.uk

[b]Pensions Institute, Cass Business School

[c]Durham University Business School

[d]Prudential Financial

1 Introduction

The field of stochastic mortality modelling has seen rapid growth in recent years, building on the early work of Lee and Carter (1992). Different strands of research have emerged in demography,[1] in statistical methodology,[2] and in the development of new models. The latter, in particular, has seen rapid progress since the early 2000's, driven by emerging risk management applications in the pensions and life insurance industries.[3] Model development has been driven by the need to have a better description of the underlying pattern of mortality improvements. Examples include: multifactor models,[4] allowance for cohort (year of birth) effects,[5] and multi-population models.[6][7]

The great majority of studies that are concerned with modelling build on the assumption that the underlying population data (typically deaths and exposures) are accurate. As a rare exception to this rule, Cairns et al. (2009) , amongst others, identified that the data for the 1886 cohort of males in the England & Wales national data were inaccurate, leading to its exclusion from their modelling work. However, they assumed that the remaining data for England & Wales males were accurate. They found some unusual patterns of mortality amongst some other birth cohorts, but they focused on finding explanations for these effects, rather than question the accuracy of the underlying data. For example, the 1918, 1919 and 1920 cohorts in England & Wales males exhibit unusual characteristics which have been attributed to a combination of lifelong *frailty* effects linked to the end of World War I and the devastating Spanish flu epidemic. Further discussion of these effects can be found in Richards (2008), who speculates, additionally, that the real cause of the 1919 cohort effects might be down to misstatements of the exposure to risk due to uneven patterns of birth. This line of reasoning is developed further in Section 2.

Two types of data are available to mortality modellers. The first consists of death counts which are typically available by calendar year and individual year of age. The format of the data varies from country to country. For example, in England & Wales, death counts, $D(t, x)$ represents the number of persons who died in year t,

[1] For example, Booth et al., 2002, and Hyndman and Ullah, 2007.

[2] For example, Brouhns et al., 2002, Czado et al., 2005 , Li et al., 2009.

[3] For example, Blake and Burrows, 2001, Blake et al., 2006, Coughlan et al., 2007, Cairns et al., 2008, Dahl et al., 2008, Li and Hardy, 2011, Li and Luo, 2012, Blake et al., 2013, Cairns et al., 2013, Cairns 2011, 2013, 2014.

[4] For example, Cairns et al., 2006, 2009, Hyndman and Ullah, 2007, Plat, 2009, Currie, 2011, Hunt and Blake, 2014, and Mavros et al., 2014.

[5] For example, Willets, 2004, Renshaw and Haberman, 2006, Cairns et al. 2009, 2011a.

[6] For example, Li and Lee, 2005, Cairns et al., 2011b, Dowd et al., 2011, Jarner and Kryger, 2011, Li and Hardy, 2011, and Börger et al., 2014.

[7] Cairns (2014) has charted the *genealogy* of these new models. He argues that the accompanying complexity might not deliver improved forecasts, and that there is a need to get back to simpler, more robust models (see, also, Mavros et al., 2014, Hunt and Blake, 2014).

aged x last birthday at the date of death. But, in other countries, deaths might be further subdivided by year of birth, or the definition of the age, x, might be different or refer to the date of registration rather than the date of death.

The second consists of exposures (a measure of the underlying population), again frequently available by calendar year and individual year of age. We define $P(s, x)$ as the number of persons aged x last birthday at exact time s. Typically this quantity is only calculated once per year rather than continuously. In England & Wales, for example, the Office for National Statistics (ONS) estimates the *mid-year* population, $P(t + \frac{1}{2}, x)$. The ONS then calculates death rates by dividing death counts by mid-year population estimates: that is,

$$m(t, x) = D(t, x)/P(t + \frac{1}{2}, x).$$

Normally, though, the death rate is defined as

$$m(t, x) \quad = \quad D(t, x)/E(t, x) \tag{1}$$

where $E(t, x)$ is called the central exposed to risk, and is equal to the mean of the population aged x last birthday during the year (see, for example, Wilmoth et al., 2007). Thus

$$E(t, x) \quad = \quad \int_0^1 P(t + s, x)ds. \tag{2}$$

England & Wales is typical of a number of countries in which population numbers are estimates based on data from a number of sources, primarily decennial censuses. Censuses are affected by uncertainty in the extent of under-enumeration and, between censuses, the estimation of net emigration[8] at individual ages is difficult.

Following the 2011 census, the ONS published revisions to the 2001-2011 intercensal population estimates (ONS, 2012a,b).[9] The extent of these revisions is highlighted in Figure 1. Death rates above age 60 are shown in the left-hand plot with only modest differences up to about age 88. Above age 88, the post-2011-censal revisions to the underlying population caused a significant increase in crude death rates. The relative magnitude of these revisions is shown in the right-hand chart in Figure 1, where we plot the ratios of the crude death rates after the revision to those before. At most ages above age 90, the increase was about 15%. Between ages 40 and 85, the impact of the revisions was modest, but below age 40, there were also significant revisions, which are most likely due to adjustments in migration. At the highest ages, migration is very modest and so is unlikely to be the main reason for the change. It is more likely that the population at ages 80+ was slightly overestimated at the 2001 census which led to bigger overestimates by 2010 (see Figure 4 below).

[8]Net emigration is defined as the number of emigrants minus the number of immigrants.

[9]There were no revisions in 2012 to population estimates before 2001.

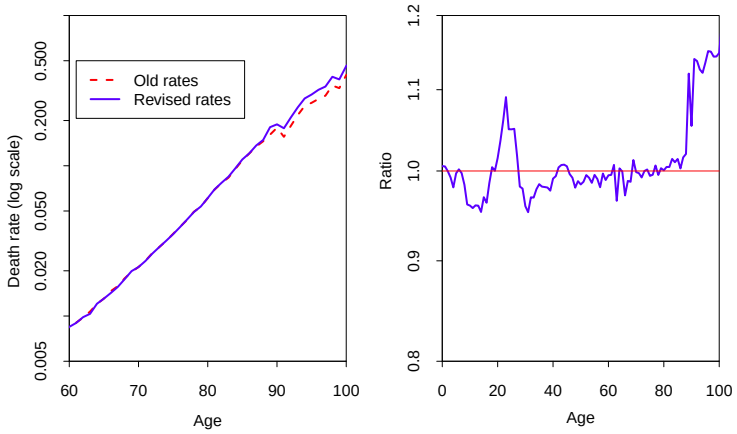

Figure 1: England & Wales males death rates in 2010. Dashed red lines: based on post-2001-censal population estimates. Solid blue line: based on 2001-2011 intercensal population estimates. Left plot: death rates on a log scale. Right plot: ratio of revised death rates to post-censal death rates in 2010.

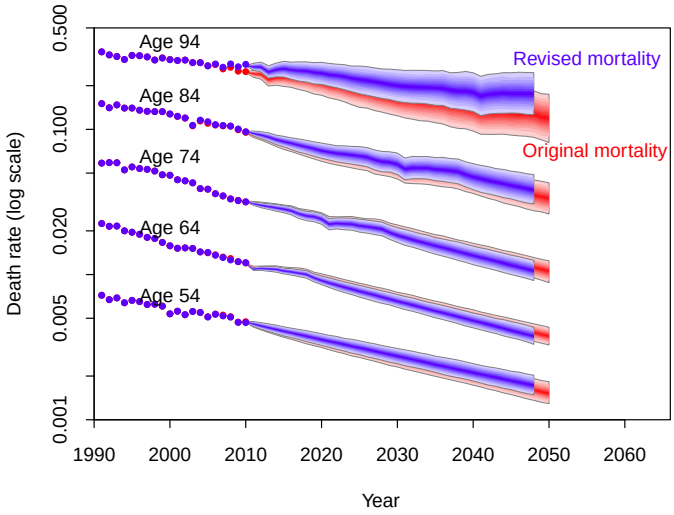

Figure 2: Historical and forecast mortality under the second-generation CBD model (M7, Cairns et al., 2009). Dots: historical death rates. Fans: forecast levels of uncertainty in death rates ranging from the 5% to 95% quantiles. Red dots and fans: historical and forecast mortality using the post-2001-censal population data. Blue dots and fans: historical and forecast mortality using the revised 2001-2010 intercensal population data.

These revisions have an impact on our predictions of future rates of mortality at particular ages. Such predictions are, necessarily, model based, but an example is shown in Figure 2 which uses the second-generation Cairns-Blake-Dowd (CBD) model, M7, proposed by Cairns et al. (2009). In general, revisions to the historical data will have three types of impact on mortality forecasts:

- the base table (most obviously at age 94);
- the central trend or improvement rate (for selected ages between 54 and 94);
- uncertainty around the central trend (not much change).

Revisions of a similar magnitude after the 2001 census can be inferred from a comparison of current ONS historical population tables with the original population data for 2000 published by the ONS (ONS 2002a).

It was these censal revisions to the population data that led to the wider study of data errors that we conduct in this paper. It is, of course, inevitable that there will be errors in both population and (to a lesser extent) deaths data, and that the magnitude of this problem will vary from country to country. What we seek to do in this study is analyse population and deaths data in as objective a manner as possible (a) to identify potential anomalies or errors in the data, and (b) to 'clean' the data prior to embarking on a mortality modelling exercise. Understanding and quantifying errors is important for a number of stakeholders with interests in: population mortality forecasts; forecasts of sub-population mortality; calibration of multi-population mortality models; assessment of levels of uncertainty in mortality forecasts; the calculation of life insurer liabilities and economic capital; pension plan buyout pricing; the assessment of basis risk in longevity hedges; and the assessment of hedges and hedging instruments.

In Section 2, we discuss in detail how errors might arise in deaths, population and exposures data, and propose a new methodology, which we call the Cohort-Births-Deaths (CBD) Exposures Methodology for calculating exposures data that exploits monthly or quarterly births data. The CBD methodology is also used to improve the calculation of mid-year population estimates based on census data. In Section 3, we develop some model-free graphical diagnostics that allow us to identify anomalies in population data. We find that the more significant anomalies can usually be linked to unusual patterns of birth rates. We propose a method in Section 4 for quantifying errors in exposures data. Although this is a Bayesian, model-based approach, we try deliberately to keep the model simple to allow the data to speak for themselves as much as possible. Section 6 looks at the impact of these error adjustments on forecasts and financial calculations. We provide further discussion and proposals for a way forward in Section 6.

2 Data Issues: Deaths, Population, Exposures

In this section, we review the reliability of the building blocks of the death rate $m(t,x) = D(t,x)/E(t,x)$. We will start with a brief discussion of death counts, before moving on to population estimates and their relationship with exposures. We will discuss how a detailed knowledge of birth patterns can help explain some of the errors that might arise in practice.

2.1 Deaths

Published death counts in England & Wales (EW), and in many other countries, $D(t,x)$, measure the number of deaths in calendar year t of people aged x last birthday at the date of death. In general, these statistics are generally accepted as accurate, but there are some issues that need to be noted.

In any given year, the total number of deaths, $\sum_x D(t,x)$, can also be regarded as accurate: it is extremely difficult for any death to go unreported in EW. But, it is possible for there to be an error in the reported age at death (e.g., the informant does not know for sure the age of the deceased). To combat this, death registrars will carry out *some* some cross checking at the point of registration where this is possible and afterwards on government databases (ONS, 2013). Cross-checking is not always possible (e.g., for immigrants) leaving open the possibility for errors to creep in and, where checks are not carried out, there might be biases towards certain ages being recorded (e.g., if the deceased was thought to be 'about' age 80, then the age at death is reported as age 80). We have assumed in the analyses that follow that these potential errors are not material in the EW data. However, we do revisit this assumption towards the end of the paper.

A second issue concerns incurred but not yet reported deaths (IBNR). Specifically, a small proportion of deaths occurring in year t will not be registered in year t (e.g. ONS, 2013). Mostly, this is because the date of death is towards the end of December and is not registered until January the following year. But a small proportion of deaths require an inquest and the death cannot be registered until the inquest has reached a verdict on the cause. This can delay registration by some months and occasionally by more than a year.

These delays influence how the ONS reports death counts. For deaths in year t, the ONS first publishes in year $t+1$ an estimate, $\tilde{D}(t,x)$, that represents deaths at age x that were *registered* in year t: this is a statistic that can be published much more quickly than deaths *occurring* in year t. Subsequently (year $t+2$), $\tilde{D}(t,x)$ is revised to $D(t,x)$: the deaths occurring in year t (see ONS, 2013). For modelling purposes, therefore, death counts in the most recent year need to be treated with some caution as there will be some overlap with the previous year (double counting) with potential biases in forecasts linked to the severity of the weather and strength

of any flu epidemic in December (which particularly affect death rates amongst the very old).

2.2 Population Estimates, Exposures, Death Rates

A potentially greater source of estimation error lies in the exposures, $E(t,x) = \int_0^1 P(t+s,x)ds$ (sometimes called the *central exposed to risk*). Since the population is not monitored or estimated on a continuous basis, it is not possible to derive the exact calculation of this expression. For example, in EW, mid-year population estimates are published annually, and the ONS then use the approximation $E(t,x) \approx P(t+\frac{1}{2},x)$.[10] [11] In contrast to the ONS, the Human Mortality Database adopts a more sophisticated approach to exposures, as documented in Wilmoth et al. (2007) (see Appendix A).

Errors in $E(t,x)$ can occur in the following ways:

- Type 1: mid-year population estimates can be inaccurate. Typical reasons for this in EW would be: no universal ID card system; censuses are 10 years apart and suffer from under-enumeration problems; and ages might be misreported at census dates.
- Type 2: the estimation of migration from year to year between censuses is difficult.
- Type 3: the assumption that $E(t,x) = P(t+\frac{1}{2},x)$ might be a poor approximation at some ages.
- Type 4: the method used to derive mid-year population estimates from census data might be inaccurate at some ages.

Errors of type 1 and 2 are well known, are difficult to reduce except at considerable cost, and lead to periodic revisions to population data, especially after decennial censuses. Errors of types 3 and 4 are less widely known and we will argue in this section that the current methodology for estimating exposures can be improved upon if we have access to monthly or quarterly births data.

[10] Some countries operate in a similar way to EW. Other countries publish estimates of population on 1 January each year, and employ the equally simple approximation $E(t,x) \approx \frac{1}{2}(P(t,x)+P(t+1,x))$.

[11] An alternative interpretation of the ONS approach is that they *define* $m(t,x) = D(t,x)/P(t+\frac{1}{2}+x)$. Either way, users of death rate data need to understand better that the numbers they are working with are not necessarily exactly what they believe.

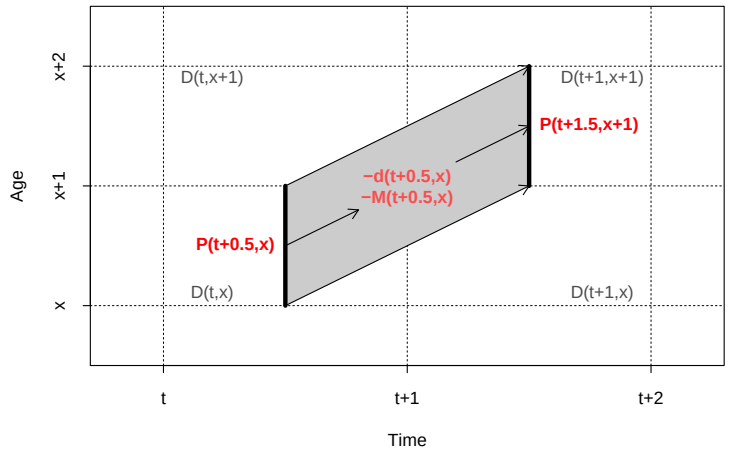

Figure 3: Relationship between mid-year population counts following cohorts and reported death counts. Population decrements include deaths and net emigration. $P(s+\frac{1}{2},y)$: mid-year population estimates. $D(s,y)$: reported death counts. For a given (t,x) cohort, the mid-year population count changes from $P(t+\frac{1}{2},x)$ to $P(t+\frac{3}{2},x+1)$ with the difference due to deaths and net emigration attributable to that cohort: $d(t+\frac{1}{2},x)$ and $M(t+\frac{1}{2},x)$ respectively.

2.2.1 Propagation of General Errors Through Time

It is clear that once errors in population counts arise, they will persist through time with the associated cohort, i.e., the errors follow the cohort.

To understand this, see Figure 3. First, assume that mid-year population counts, deaths and migration are accurate. Then, we have the identity

$$P(t+\frac{3}{2},x+1) = P(t+\frac{1}{2},x) - d(t+\frac{1}{2},x) - M(t+\frac{1}{2},x)$$

where $d(t+\frac{1}{2},x)$ is the number of individuals in the cohort who die between $t+\frac{1}{2}$ and $t+\frac{3}{2}$, and $M(t+\frac{1}{2},x)$ is the corresponding net emigration *from* the cohort.

If, in fact, there is an error in the population count, but accurate deaths and net emigration figures in the following year, then we have the identity

$$\hat{P}(t+\frac{3}{2},x+1) = \hat{P}(t+\frac{1}{2},x) - d(t+\frac{1}{2},x) - M(t+\frac{1}{2},x)$$

where $\hat{P}(s+\frac{1}{2},y) = P(s+\frac{1}{2},y) + \epsilon_P(s+\frac{1}{2},y)$ and $\epsilon_P(s+\frac{1}{2},y)$ is the error in the population count. If deaths and net emigration figures are accurate, then it is straightforward to see that $\epsilon_P(t+\frac{1}{2},x) = \epsilon_P(t+\frac{3}{2},x+1)$: that is, the error at time $t+\frac{1}{2}$ follows the cohort. Finally, death counts and (much more so) net emigration

are themselves subject to errors and so

$$\epsilon_P(t+\frac{3}{2},x+1) = \epsilon_P(t+\frac{1}{2},x) + d(t+\frac{1}{2},x) - \hat{d}(t+\frac{1}{2},x) + M(t+\frac{1}{2},x) - \hat{M}(t+\frac{1}{2},x).$$

At younger ages, errors in net emigration figures from year to year will tend to dominate, but, at higher ages, the error in the population count at time $t+\frac{1}{2}$ will dominate until such time as the population counts are subject to revision (e.g., at the next census). We call the overestimated population "phantoms". This is because phantoms never die (at least between period revisions). That is, if we overestimate the population ($\hat{P}(t+\frac{1}{2},x) > P(t+\frac{1}{2},x)$), then the $\epsilon_P(t+\frac{1}{2},x)$ phantoms persist through time, while members of the true population die off. Figure 4 illustrates this in a stylised way. The cohort has been overestimated by 200 at the time of the census at time 0. The extra 200 form a growing percentage of the total over the next 10 years until the time of the next census. Additionally, the reported death rate (*actual* deaths divided by *over-estimated* population) will be lower than the true death rate, and the relative error in the reported death rate will grow over the 10 years.

After 10 years (Figure 4), there is a further census and it is estimated that the cohort is rather smaller (dark grey bar) than the time-10 post-censal estimate (black and light grey bars), although the revised estimate is likely still to differ from the true value. The question then arises: how do we accommodate this change in population estimates at time 10? What the ONS does is, rather than recognise that the error was made at the time of the previous census, it assumes that its previous population estimate was accurate and linearly "kills off" the phantoms. If the required adjustment is A at time 10, then the ONS adopts the simple approach of deducting $A/10$ from the cohort population in each of years 1 to 10 in addition to the existing post-censal decrements (ONS, 2002c, Duncan et al., 2002) (see the red down arrows in Figure 5). This means that following the new census, population estimates for the previous 10 years are likely to be revised. Specifically, let $P_P(s+\frac{1}{2},x+s)$ be the post-censal estimates and $P_I(s+\frac{1}{2},x+s)$ the inter-censal estimates. The adjustment at the time 10 census is $A = P_P(10\frac{1}{2},x+10) - P_I(10\frac{1}{2},x+10)$ and we then define $P_I(s+\frac{1}{2},x+s) = P_P(s+\frac{1}{2},x+s) - A \times s/10$ for $s = 1,\ldots,10$.

Implicit within this procedure is the assumption that the census measure at time 0 was correct. But the stylised example presented in Figures 4 and 5 suggests that, in some circumstances, the ONS's method only partly fixes the problem. Furthermore, the stylised example is more likely to be representative of reality (if exaggerated) for the oldest ages than at younger ages. At younger ages, it is very likely that errors at time 10 are simply due to the difficulties in estimating net emigration.

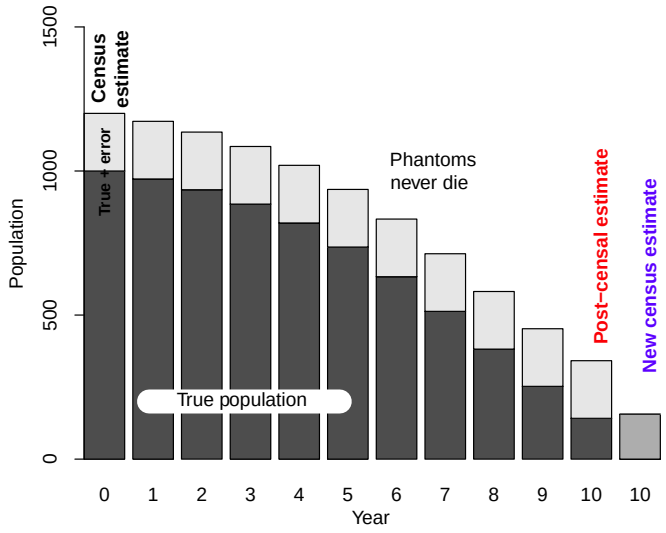

Figure 4: Stylised representation of the changing cohort size between censuses at times 0 and 10. Black bars: true cohort size. Light grey bars: additional phantoms. Dark grey bar (time 10): revised cohort size at the time 10 census.

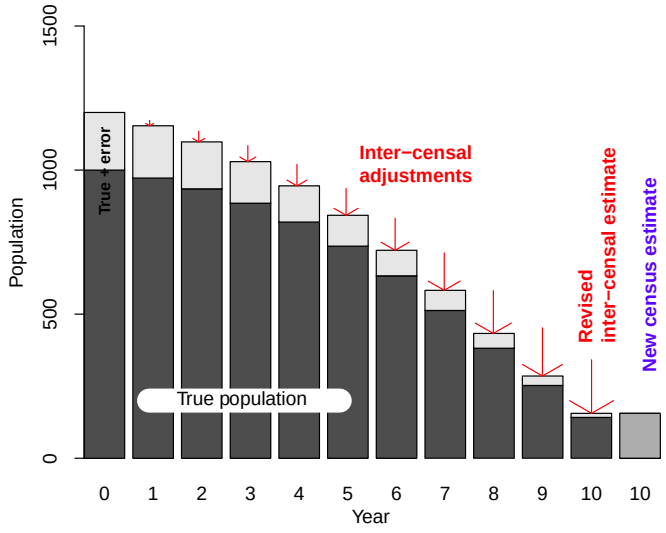

Figure 5: Inter-censal adjustments to the cohort size. Red arrows: size of change from post-censal population estimates to inter-censal estimates. Revised intercensal estimate equals the new, time-10 census estimate.

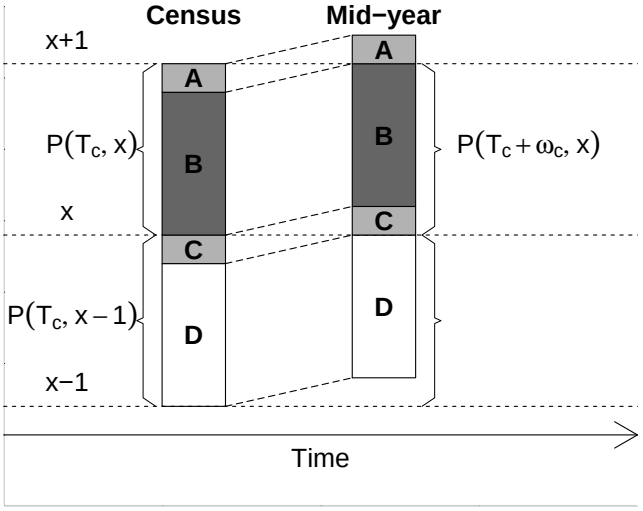

Figure 6: Population progression between a census date, T_c, and the middle of the year, $T_c + \omega_c$. People in group A are aged x at time T_c and have a birthday between T_c and $T_c + \omega_c$ and advance to age $x + 1$. B: age x at both T_c and $T_c + \omega_c$. C: age $x - 1$ progressing to x. D: $x - 1$ at both dates.

2.2.2 Census to Mid-year Shift

The next potential error concerns the census to mid-year shift. In EW, censuses occur every 10 years on variable dates in the spring.[12] The issue is illustrated in Figure 6. Suppose the census occurs at time T_c and the middle of the same year is, say, $T_c + \omega_c$. At $T_c + \omega_c$, we wish to estimate the population aged x last birthday. This will consist of a mixture of people aged x and $x - 1$ last birthday at the time of the census. $P(T_c + \omega_c, x)$ will consist of those aged x at time T_c who did not have a birthday between T_c and $T_c + \omega_c$ (Figure 6, group B), plus those aged $x - 1$ at time T_c who did have a birthday between the census and the middle of the year (group C). $P(T_c + \omega_c, x)$, therefore, consists of appropriate proportions of $P(T_c+, x)$ and $P(T_c, x - 1)$ (Figure 6), adjusted for deaths and net emigration between T_c and $T_c + \omega_c$.

In 2001, the ONS used the assumption that birthdays were spread evenly throughout the year at all ages (Duncan et al., 2002). On this basis, a proportion ω_c of those aged $x - 1$ at the census would reach their x^{th} birthday before 30 June 2001, and a proportion $1 - \omega_c$ aged x would remain aged x.[13] As a consequence

$$P(T_c + \omega_c, x) = \omega_c P(T_c, x - 1) + (1 - \omega_c)P(T_c, x) - \text{deaths} - \text{net emigration}.$$

We have not been able to establish what assumptions were employed in other census

[12] Most recently: 25/4/1971; 5/4/1981; 21/4/1991; 29/4/2001; 27/3/2011. In the US, censuses fall every 10 years on 1 April; most recently 1/4/2010.

[13] The census was on 29 April 2001, so $\omega_c = 62/365$.

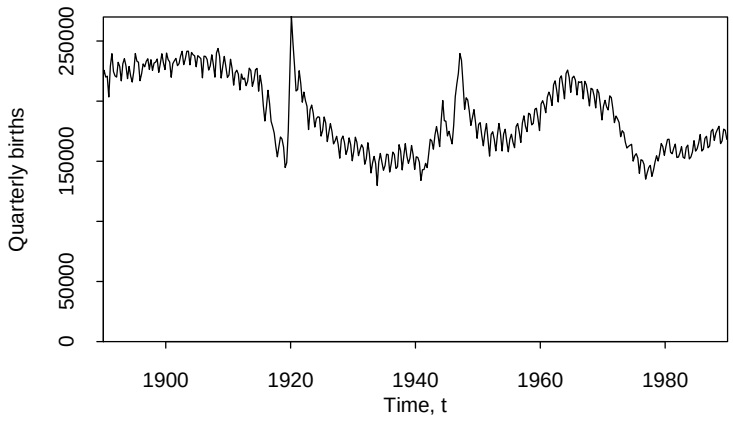

Figure 7: EW quarterly births from 1890 to 1990.

years. However, on the basis of statistical analysis, we conjecture in Section 3 that different assumptions were probably employed.

The assumption of an even distribution of births is reasonable for most birth cohorts when birth rates were reasonably stable from year to year. However, for a few cohorts the assumption results in a very poor approximation.

2.2.3 The Cohort-Births-Deaths (CBD) Exposures Methodology

We now propose an alternative approach that refines the even distribution of births assumption. Our underlying hypothesis is simple: *at any point in time, t, the pattern of birthdays at age x will reflect the actual pattern of births x years earlier.* At older ages, the relative proportions across a single age needs to reflect mortality. The accuracy of the assumption will depend on the proportion of the population at a given age that are immigrants (born and registered outside EW) and on any differences between the average birth pattern of immigrants and those born in EW.

In Figure 7, we have plotted quarterly births from 1890 to 1990. Two features are apparent: a clear seasonal pattern that persists to recent times; and significant fluctuations in birth rates over time. In particular, sudden jumps in birth rates after the first and second world wars, and we argue below that these jumps can have an impact on how mid-year population estimates should be derived from census data.

We can use the births data to estimate what proportion of those aged $x-1$ at the time of the census will reach their next birthday before 30 June, and what proportion of those aged x at the census will still be x on 30 June. This can be used to calculate prospectively: *out of those aged $x-1$ and x at the time of the census, how many are due to be (i.e., ignoring deaths) aged x last birthday on 30 June*? We then compare

these CBD-based numbers for each age x with the uniform distribution of birthdays used by the ONS in 2001.

To simplify the discussion very slightly let us assume that the 2001 census took place right at the end of April, to allow us to focus on months of birth only. As an example, let us look at males born between May 1918 and April 1920 who were thus aged 82 or 81 at the time of the 2001 census. Tables 1 and 2 show how the CBD and ONS methods, respectively, work in terms of estimating how many males will be aged 82 on 30 June 2001 based on those aged 81 and 82 at the census date. This calculation for age 82 reveals the most significant difference between the CBD and ONS methodologies.

The figures 72741 (CBD) and 79352 (ONS) make the assumption that there are no deaths or net emigration between the census and 30 June. The ONS published mid-year population at age 82 in 2001 of 78615 does adjust for deaths and net emigration between the census and 30 June, and so, under the CBD method, we would propose to multiply the ONS mid-year population estimate of 78615 by 72741/79352, giving a revised population estimate of 72065 for age 82 on 30 June 2001.[14]

The difference between the CBD and ONS uniform distribution of deaths assumptions in the 2001 mid-year population estimates is illustrated in Figure 8 (red dashed line). The 1919 cohort discussed above stands out, but the 1920 cohort also has a large difference and there are also significant differences between the methods for cohorts born during and after the second world war (there was a sharp baby boom in 1947; see Figure 7). At most other ages, the difference is at most $\pm 1\%$.[15]

The CBD method can be refined further by adjusting the number of births in Table 1 for mortality between the exact date of birth and the current measurement date. This would have the impact of reducing slightly the *proportion* of 82-year-olds at the census date who were born in May-June 1918. This is illustrated in Figure 8 (thin grey line) using a stylised Gompertz pattern of mortality. It can be seen that adjusting for mortality has little effect at younger ages, and only just starts to become noticeable amongst the oldest cohorts plotted.

We conclude that the CBD method is robust relative to the mortality assumption. However, as already remarked, the CBD method makes no allowance for migration. The approach might be less accurate if immigrants form a significant proportion of the population at certain ages, and if the pattern of births for these immigrants is significantly different from EW births.

[14]Equivalently, we could take the unadjusted CBD figure (72741) and multiply this by the ratio of the ONS mid-year population estimate at age 82 in 2001 adjusted for deaths and net emigration to the corresponding ONS unadjusted population estimate (78615/79352) to give the same revised population estimate of 72065 for age 82 on 30 June 2001.

[15]However, at high ages, e.g. age 80, an error of 1% in a census year can grow to three times that over the next 10 years (see Figure 4), so even these modest errors are potentially important and deserve our attention.

Birth month	No. of births	Age on 30/4/2001	Proportion	2001 census	CBD estimate	Age at mid-year	CBD mid-year
5-6/1918	113475	82	0.17785	72114	12825	83	72741
7/1918-4/1919	524566	82	0.82215		59289	82	
5-6/1919	99174	81	0.11642	115545	13452	82	
7/1919-4/1920	752725	81	0.88358		102093	81	

Table 1: Calculation of mid-year population estimates based on population at the end-April-2001 census date using the CBD assumption, and assuming no deaths or net emigration between the census and 30 June. Birth counts are for males and females. 2001 census counts from ONS (2002b).

Birth month	Age on 30/4/2001	Proportion	2001 census	ONS estimate	Age at mid-year	ONS mid-year
May-June 1918	82	2/12	72114	12019	83	79352
July 1918-April 1919	82	10/12		60095	82	
May-June 1919	81	2/12	115545	19257	82	
July 1919-April 1920	81	10/12		96288	81	

Table 2: Calculation of mid-year population estimates based on population at the end-April-2001 census date using the ONS uniform distribution of birthdays assumption, and assuming no deaths or net emigration between the census and 30 June.

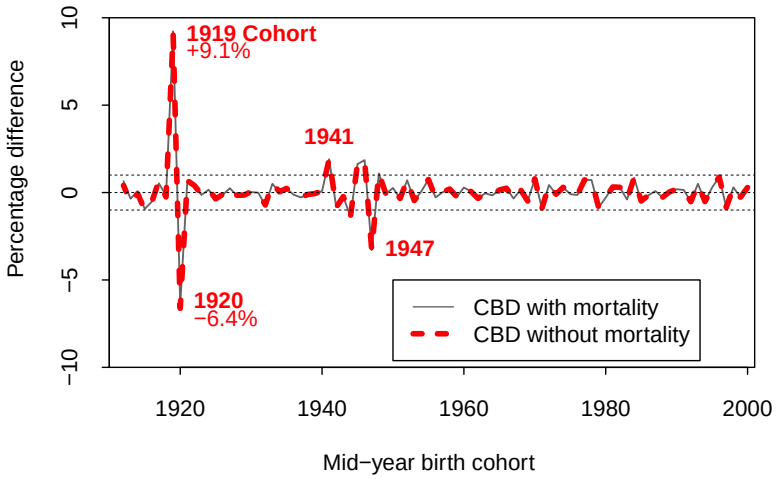

Figure 8: Estimated relative difference between mid-year population estimates by year of birth, for the ONS uniform distribution of births assumption used in 2001, and the CBD method. Solid grey line: ONS versus CBD with a stylised Gompertz mortality assumption and no migration. Dashed red line: ONS versus CBD with no mortality or migration.

2.2.4 $P(t+\frac{1}{2},x) \approx E(t,x)$: How Good is this Approximation?

The true exposures are $E(t,x) = \int_0^1 P(t+s,x)ds$ (equation 2), but, since the ONS publishes only mid-year population estimates, it (in common with many other agencies) makes the assumption that $E(t,x) = P(t+\frac{1}{2},x)$ leading to $m(t,x) = D(t,x)/P(t+\frac{1}{2},x)$. So how good is this approximation? The answer lies in the convexity of the underlying population function as illustrated in Figure 9. As can be seen, the approximation will be quite good if the function $P(t+s,x)$ is reasonably linear in the interval $0 < s < 1$. However, if $P(t+s,x)$ is non-linear (right hand plot) then there could be a significant difference between $P(t+\frac{1}{2},x)$ and the true $E(t,x)$.

We can investigate readily how this approximation works in practice by using births data to estimate the population aged 0 at any point in time. Let $\tilde{P}(t,0)$ represent the population at time t aged 0 last birthday, assuming no deaths and net emigration. Although, infant mortality was high in the early 20th century (approximately 1 in 10 died in their first year, mostly very soon after birth), the pattern of mortality changes much more slowly over time relative to the variation in birth rates that we seek to analyse here. Additionally, we only make use of age 0 exposures *relative* to the mid-year population counts. The resulting ratios will be relatively insensitive to the underlying rates of infant mortality, provided these change slowly over time.

$\tilde{P}(t,0)$ can be calculated quarterly or monthly (depending on the granularity of

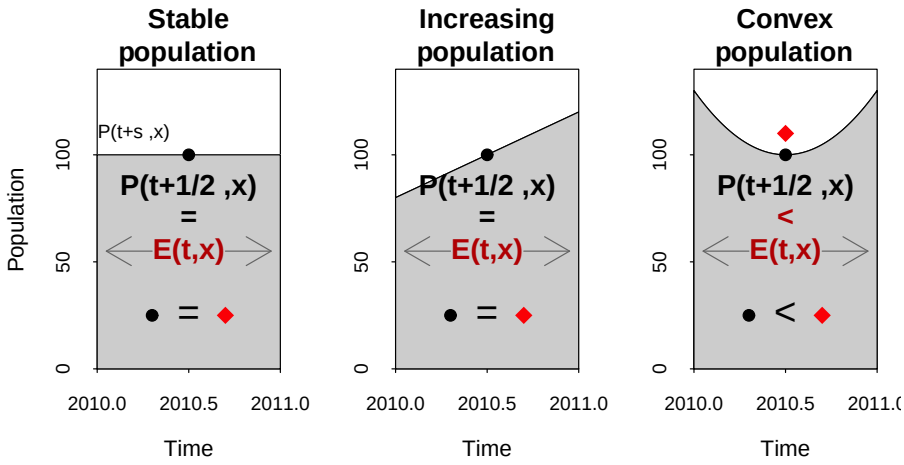

Figure 9: Stylised representation of the relationship between the population function $P(t+s,x)$ and the exposures $E(t,x)$. The true $E(t,x)$ equals the shaded area under the curve (red diamond) and this is to be compared to $P(t+\frac{1}{2},x)$ (black dot).

births data) and is, simply, the sum of the births in the preceding 12 months and is shown over the period 1910 to 1925 in Figure 10 (continuous black line). The mid-year population estimates at age 0 assuming no deaths or net emigration are shown as black dots. The equivalent exposure estimates are $\tilde{E}(t,0) = \int_0^1 \tilde{P}(t+s,0)ds$ and these are approximated using Simpson's rule[16] based on the quarterly $\tilde{P}(t+s,0)$. These exposure estimates are shown in Figure 10. Mostly the $\tilde{E}(t,0)$ (red diamonds) are very close to the mid-year $\tilde{P}(t+\frac{1}{2},0)$ (black dots). However, where there is significant non-linearity, differences can be quite striking. And the most significant non-linearities and differences occur when there is a sudden baby boom,e.g., at the end of 1919.

Irregular patterns of birth convert into similar patterns of birthdays in later years and so we argue that the relationship between the mid-year population estimate at age x and exposures should be approximately the same as it was at age 0 for the same cohort. We, therefore, propose a convexity adjustment ratio to be applied to mid-year population estimates, $P(t+\frac{1}{2},x)$, namely

$$E(t,x) = P(t+\frac{1}{2},x) \times \frac{\tilde{E}(t-x,0)}{\tilde{P}(t+\frac{1}{2}-x,0)} \tag{3}$$

where we call $CAR(t-x) = \tilde{E}(t-x,0)/\tilde{P}(t+\frac{1}{2}-x,0)$ the Convexity Adjustment Ratio (CAR) for the $t-x$ birth cohort. The CAR is plotted in Figure 11. As we can see, the CAR is generally close to 1 (that is, the assumption that $E(t,x) = P(t+\frac{1}{2},x)$ is a reasonable approximation). But in a few years (1919, 1920, 1946, 1947) there is a significant deviation from 1. And these deviations coincide with periods when

[16]See Appendix A.

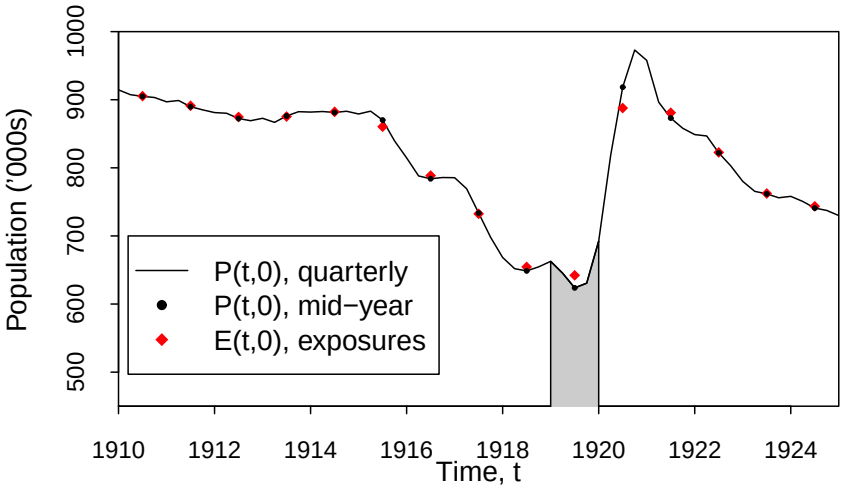

Figure 10: Relationship between age 0 population and exposures estimates. Solid line: quarterly $\tilde{P}(t,0)$. Black dots: mid-year $\tilde{P}(t+\frac{1}{2},0)$. Red diamonds: exposures $\tilde{E}(t,x)$.

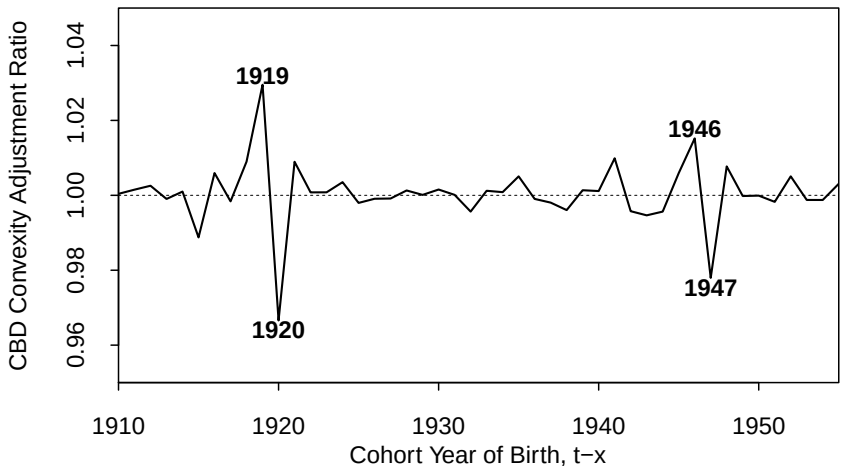

Figure 11: The convexity adjustment ratio, $CAR(t-x) = \tilde{E}(t-x,0)/\tilde{P}(t+\frac{1}{2}-x,0)$, as a function of year of birth $t-x$.

there were rapid changes in birth rates around the end of the first and second world wars.

2.2.5 Application of the CBD Exposures Methodology

Summarising Sections 2.2.3 and 2.2.4, we would propose two actions. First, mid-year population estimates can be improved in census years using the CBD exposures methodology, augmented by existing ONS procedures for following cohorts between censuses to allow for deaths and migration. We have not attempted to apply this method retrospectively as we do not have accurate accounts of how mid-year population estimates were derived in census years other than 2001. However, backfilling methods could be used to interpolate between census years after these have been corrected. Second, and in addition to these census-to-mid-year adjustments, exposures can be calculated using the convevity adjustment ratio.

2.2.6 High Age Methodology

The ONS mid-year population estimates give population counts at individual ages up to age 89 and then a single number each for males and females for age 90 and above. Death counts are available for individual ages above 90. The ONS use the Kannisto-Thatcher (KT) method (an adaptation of the extinct cohort method; see, Thatcher et al., 2002) to calculate population estimates for individual ages up to age 104 (ONS, 2014).

The KT method attempts to estimate exposures at individual ages from limited underlying data and, therefore, gives rise to the possibility of errors. In particular, we will investigate in the following sections the possibility that there are discontinuity errors at the age 89-90 boundary where the high-age methodology kicks in.

3 Graphical Diagnostics and Signature Plots

In this section, we outline and illustrate a variety of graphical diagnostics that are designed to help verify the quality of the data or identify where anomalies might lie. The use of a graphical diagnostic might contain the following elements:

- a hypothesis that concerns some characteristics of the underlying data;
- a specified way of plotting some aspect of the data;
- if the hypothesis is true, then the plot should exhibit a recognisable structure or pattern;

- if the plot does not exhibit the expected structure, then the type of deviation might provide us with pointers on how the underlying hypothesis might be varied or point to particular types of problem with the data.

3.1 Graphical Diagnostic 1

Hypothesis: *Crude death rates by age for successive cohorts should look similar.*

The diagnostic that we will employ to test visually this hypothesis is to plot the curve of crude death rates for each cohort against age. Since we wish to compare cohorts (as required in the hypothesis), we compare, for each cohort, the curve of death rates with its four nearest neighbours.

Our first examples are presented in Figure 12 centred on the 1909 and 1929 birth cohorts. Both the left (1909) and right-hand (1929) plots exhibit the expected pattern: namely that each of the five lines in each plot all follow a consistent upwards trend that is close to linear. Each curve exhibits a certain amount of randomness as age increases. This reflects (a) random variation in the number of deaths given the true underlying death rate (known as Poisson type risk), and (b) randomness in the period effects that drive improvements in the underlying death rates. At the higher ages, the Poisson randomness is relatively modest, given the size of the underlying population.

In the left-hand plot of Figure 12, the lines almost lie on top of each other across the full range of ages. In the right-hand plot, the lines are more or less parallel, but the positions of the five curves reflects a greater rate of improvement in mortality rates between these cohorts. This greater rate of improvement between cohorts born around 1930 is well known (see, for example, Willets, 2004). These plots are typical of most birth cohorts.

Now contrast Figure 12 with Figure 13 (left), centred on the 1919 cohort. As a graphical diagnostic, this latter plot does not exhibit the characteristics that we would expect to see (as it does in Figure 12). Instead, we can see that the solid black curve for the 1919 cohort after about age 75 trends below its neighbours. We will discuss this in more detail later, but will just remark here that the pattern is consistent with the emergence of phantoms in the 1919 cohort exposures over the period between the censuses of 1991 (when the 1919 cohort was aged 72) and 2001: in other words, the problem with the 1919 cohort began in 1991. Thereafter, the existence of the phantoms (who never die) causes the curve to drift lower still relative to its neighbours (see Figure 4).

In Figure 13 (right), we have also plotted the cohort curves of death rates after some model-based adjustments. Specifically, the crude death rates had systematic period (calendar year) effects estimated by the model M9 outlined in Appendix C. The procedure does not eliminate any of the Poisson-type randomess, so the greater

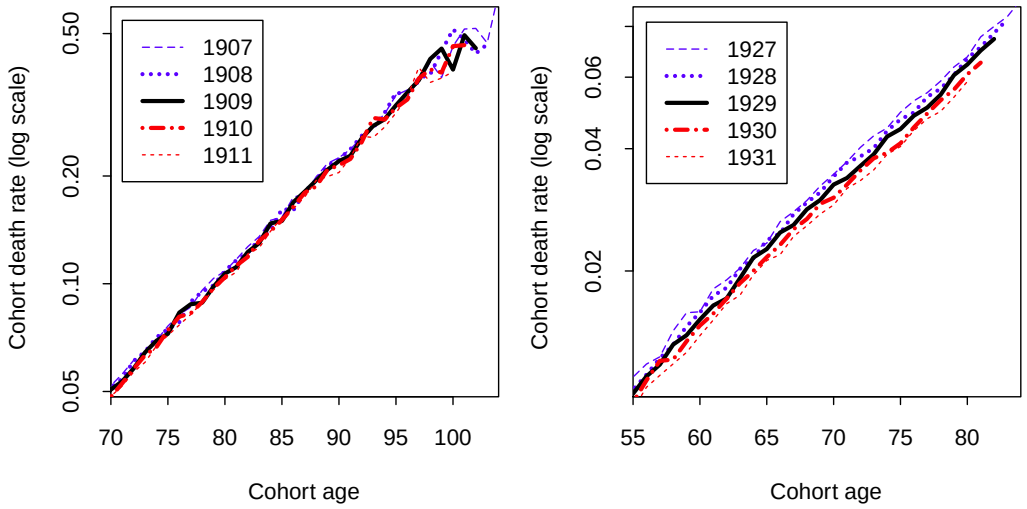

Figure 12: Cohort death rates by age for the 1907 to 1911 cohorts (left) and the 1927 to 2031 cohorts (right). The right-hand end point for each curve corresponds to 2011. ONS (revised) EW males data up to 2011.

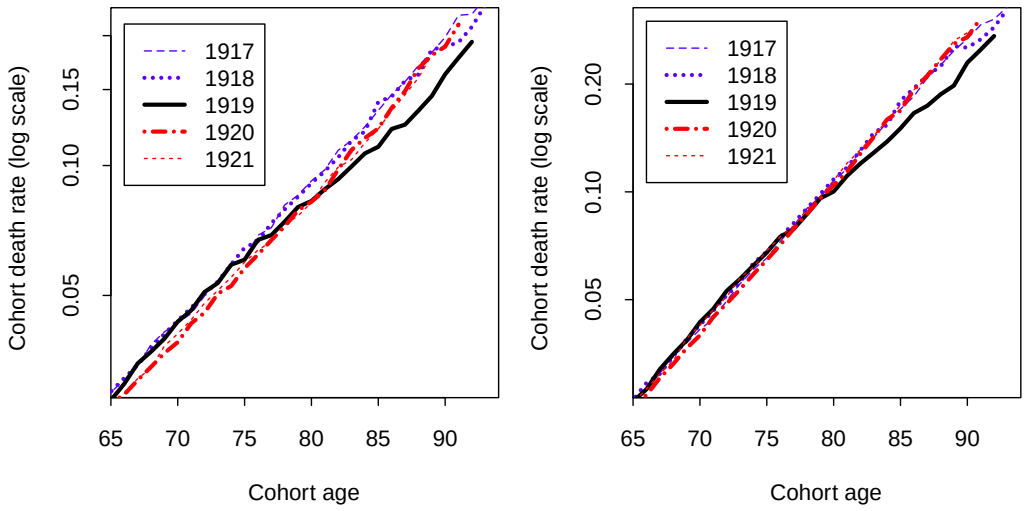

Figure 13: Cohort death rates by age for the 1917 to 1921 cohorts. Left: curves based on crude death rates. Right: modified death rates with model-based period effects subtracted out. Other details as in Figure 12.

degree of smoothness clearly indicates that randomness in the year-to-year period effects is quite significant. Although the procedure produces smoother curves, our conclusions about the 1919 cohort are clearly unchanged. Nevertheless, we mention the procedure here because it might make detection of more modest anomalies easier, albeit at the cost of introducing some modelling assumptions.

3.2 Graphical Diagnostics 2A and 2B

Hypothesis: *Underlying log death rates are approximately linear within each calendar year*

This hypothesis leads us to the use of two graphical diagnostics that allow us to visualise the same derived statistics in different ways.

As a starting point, we calculate how far individual groups of observations (three consecutive ages within a calendar year) deviate from linearity as measured using the empirical concavity function

$$C(t,x) \;=\; \log m(t,x+t) - \frac{1}{2}\Big(\log m(t,x+t-1) + \log m(t,x+t+1)\Big)$$

We then plot this in two ways:

- 2A: plot $C(t,x)$ by cohort: that is, $(x_0+s, C(t_0+s, x_0+s))$ for the (t_0-x_0) birth cohort;
- 2B: a 2-dimensional heat map of $C(t,x)$ by age and calendar year.

If the hypothesis is true, the concavity function should stay close to zero without exhibiting any systematic bias above or below.

Some typical results for graphical diagnostic 2A are plotted in Figure 14. The pattern exhibited in the plot for the 1924 cohort is consistent with the hypothesis: the dots are randomly above and below zero. However, the plots for the 1920, 1919, and 1947 cohorts all contain structure that is not consistent with the hypothesis. So we consider next what the irregular patterns tell us about the underlying data.

The 1919 and 1920 cohort plots each contain two specific effects. Before 1991, the plots are level but are consistently above (1919) or below (1920) the zero line. After 1992, both plots suddenly start to drift at a steady rate. The shift up or down is consistent with our earlier discussion in Section 2.2.4 where we argued that mid-year population estimates were not always good approximations to the true exposures for some cohorts. In the 1919 and 1920 cohort plots, we have added a horizontal dashed line which makes an allowance for this effect. The adjustment means that, up to 1991, the plots of the empirical concavity function now look more reasonable. However, the positioning of the dashed line in both cases could be

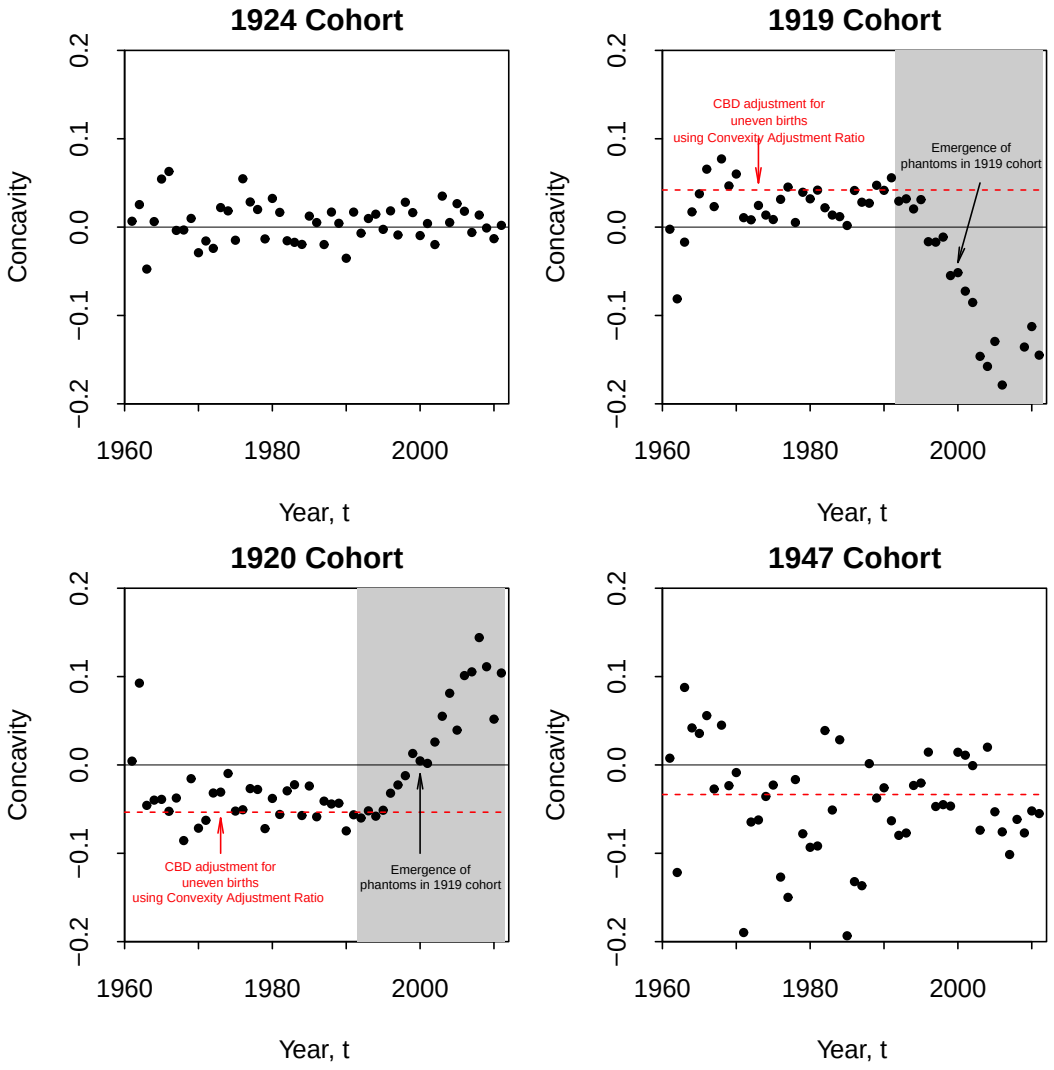

Figure 14: The empirical concavity function, $C(t+s, x+s)$, plotted against calendar year $t+s$ for birth cohorts $t-x$ =1924, 1919, 1920 and 1947. Shaded region covers the calendar years 1992 to 2011.

improved suggesting that other factors are at play beyond the proposed Convexity Adjustment Ratio. However, up to 1991, both plots are consistent with a systematic bias in the estimate of cohort exposures.

After 1992, both plots drift steadily down or up. This is consistent with the introduction of phantoms into the 1919 cohort and 'negative' phantoms into the 1920 cohort from 1992 onwards. We infer from this that the method for shifting from a census to the mid-year population estimate discussed in Section 2.2.2 was applied at the time of the 2001 census but not in previous census years. Furthermore, the apparent inconsistencies revealed in 2001 as a result of applying this method then led the ONS to backfill by cohort over the period 1992 to 2001 to provide a smoother progression of population estimates between the 1991 and 2001 censuses. In other words, the ONS has corrected an error by smoothing it out rather than identifying the cause of the error (see Figure 5). This will be discussed further when we consider graphical diagnostic 3 below.

The plot for the 1947 cohort is generally more dispersed. The reason for this is that the cohort is younger and, therefore, the underlying death counts are much smaller and more random. However, we can still see that the dots are biased below zero, to an extent that is consistent with a requirement to apply the Convexity Adjustment Ratio as indicated by the horizontal dashed line.

We move on next to graphical diagnostic 2B, and plot the empirical concavity function in two dimensions in the form of a heat map in Figure 15. Apart from some obvious structural features, we can also see greater randomness at the high ages and low ages caused by low death counts. We can normalise the empirical concavity function by dividing by an estimated standard error that is based on the Poisson assumption. The normalised concavity function, $C_N(t, x)$, is plotted in Figure 16.

If the hypothesis about linearity in the log deaths rates were true, then the heat map should essentially be completely random apart from some autocorrelation between cells within each calendar year as a result of the common contributors to, for example, $C(t, x)$ and $C(t, x + 1)$. Figure 16 does contain plenty of randomness, but it also contains significant structure. The clear diagonals reflect the cohort biases that we have already dicussed for the 1919, 1920 and 1947 cohorts as well as one or two others. But other structural elements are ones that are new. Specifically, there is a number of clear horizontal traces. In most cases, these horizontal bands are consistent with small biases in the reporting of age at death. In general, these bands of blue colour tend to be stronger towards the left suggesting that the reporting of age at death has become steadily more accurate as time progresses.

The colour bands around age 90, suggest that deaths reported at age 90 and 91 are too *low* rather than too high. Furthermore, this band is consistent across all years. An alternative explanation to bias in the reporting of age at death is based on the fact that exposures at age 90 and above are calculated using the Kannisto-Thatcher method instead of the standard census and inter-censal methods estimates used at

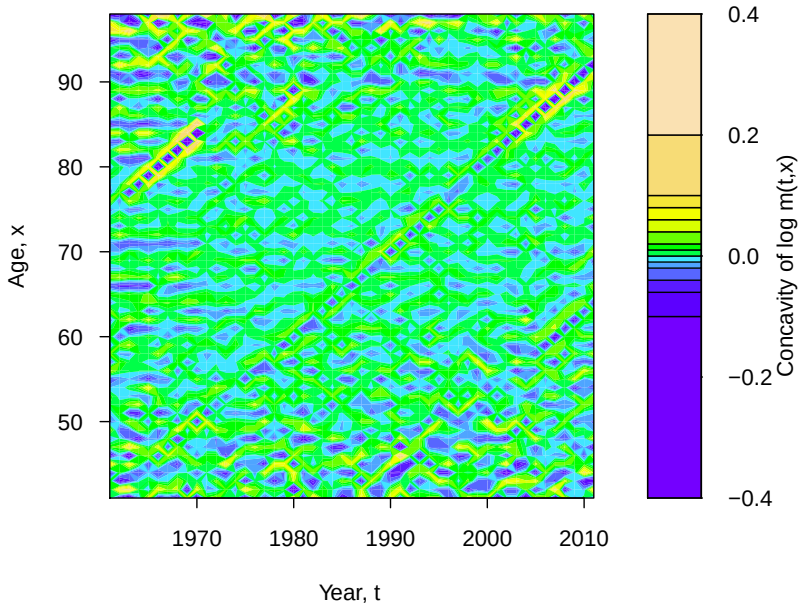

Figure 15: The empirical concavity function $C(t, x)$ for different years, t, and ages x.

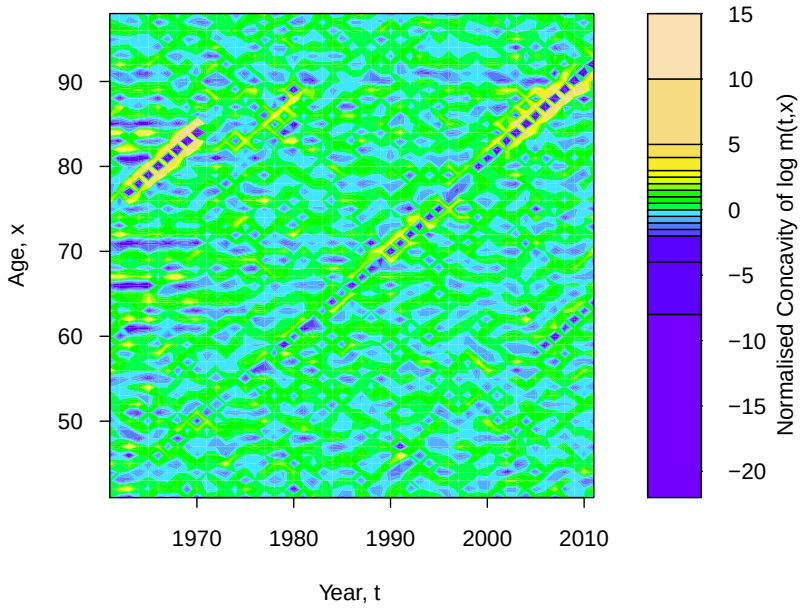

Figure 16: The normalised concavity function $C_N(t, x) = C(t, x)/S.E.(C(t, x))$. Standard errors are calculated on the assumption that deaths follow the conditionally independent Poisson assumption.

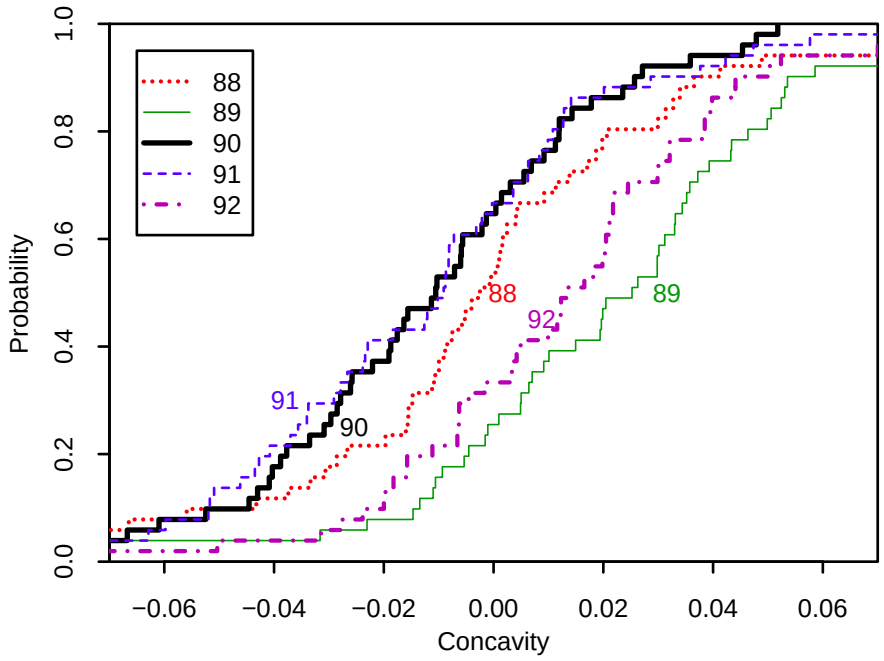

Figure 17: The empirical cumulative distribution function of the concavity function, $C(t, x)$, for ages 88 to 92.

all ages up to 89. This simple graphical diagnostic suggests that the current way in which the Kannisto-Thatcher method is employed potentially introduces a small but systematic bias.

These age-related biases are investigated further in Figure 17 where we plot the empirical cumulative distribution functions of the $C(t, x)$ for specific ages, x. The curve for age 88 looks about right (median close to 0). However, visually, the curves for ages 90 and 91 look to be well to the left and those for ages 89 and 92 look to be well to the right.[17] The negative values for ages 90 and 91 point to either significant under-reporting of deaths at ages 90 and 91 (implausible) or significant overestimation of mid-year populations at those ages (and, possibly, also age 92) by as much as 5%. Comparable plots for lower ages reveal shifts to the left and right, but, typically, the sizes of the shifts are smaller and potentially less significant. Additionally, as remarked previously, at lower ages, the age effects seem to diminish gradually with time.

[17]The significance of these biases was confirmed informally using an adapted Kolmogorov-Smirnov test that allowed for cross correlation between ages. The distribution of the test statistic under the null hypothesis (no difference) was generated using simulation, revealing that the CDF's for ages 90 and 91 were very significantly different from those for ages 89 and 92.

3.3 Graphical Diagnostic 3

Hypothesis: *Changes in cohort population sizes should match closely the pattern of reported deaths.*

Recall that we have two types of data: death counts, $D(t,x)$, and mid-year population estimates, $P(t+\frac{1}{2},x)$. We will now define the decrement

$$\hat{d}(t+\frac{1}{2},x) = P(t+\frac{1}{2},x) - P(t+\frac{3}{2},x+1)$$

which equals the sum of the deaths from the cohort aged x last birthday at time $t+\frac{1}{2}$, denoted by $d(t+\frac{1}{2},x)$, and net emigration, $M(t+\frac{1}{2},x)$ (see Figure 3). At high ages, net emigration will be modest compared to deaths and so $\hat{d}(t+\frac{1}{2},x)$ will be approximately equal to $d(t+\frac{1}{2},x)$. It is natural, therefore, to compare $\hat{d}(t+\frac{1}{2},x)$ with the reported death counts. However, as can be seen from Figure 3, four different reported death counts are linked to the decrement, $\hat{d}(t+\frac{1}{2},x)$ (namely, $D(t,x)$, $D(t,x+1)$, $D(t+1,x)$ and $D(t+1,x+1)$). Consequently, we plot $\hat{d}(t+\frac{1}{2},x)$ for each cohort, and compare this with the reported deaths by cohort. We need to make two types of adjustment to the data. First, death counts $D(t,x+1)$ and $D(t+1,x)$ will be disproportionately high or low if the underlying birth rates for these two cohorts were much higher or lower, respectively, relative to the $D(t,x)$ and $D(t+1,x+1)$ cohort. Thus deaths in adjacent cohorts are scaled to reflect this. Second, for the cohort in question, the mid-year population might not be an adequate approximation to the true exposures, $E(t,x)$, that corresponds to the deaths, $D(t,x)$. Thus, we scale the mid-year populations by the Convexity Adjustment Ratio (CAR) and, hence, the $\hat{d}(t+\frac{1}{2},x)$, before these are plotted.

We summarise the above procedure as follows. For a plot that follows the $t-x$ birth cohort:

- plot (solid dots) $\hat{d}(t+\frac{1}{2},x)\times CAR(t-x)$, where $CAR(t-x) = E(t-x,0)/P(t+\frac{1}{2}-x,0)$;
- plot (black and blue lines respectively) $D(t,x)$ and $D(t+1,x+1)$ with no adjustment;[18]
- plot (green line) $D(t,x+1) \times E(t-x,0)/E(t-x-1,0)$;
- plot (red line) $D(t+1,x) \times E(t-x,0)/E(t+1-x,0)$.

Figures 18 and 19 show typical examples of how the plot should look if the hypothesis is true. Specifically, the black dots and the four lines for the reported deaths follow each other quite closely. The four curves of reported deaths all follow each other quite

[18]The blue line is simply the black line shifted by a year.

closely, indicating that the adjustment for unequal births is about right: without this adjustment, some curves would be significantly higher or lower than the central case. The similar shape for each curve links to the hypothesis underpinning Graphical Diagnostic 1. At younger ages and earlier years, the decrements $\hat{d}(t + \frac{1}{2}, x)$ are rather more variable, reflecting year to year variations in net emigration.

Closer inspection reveals some unusual effects that indicate that there are issues with the population data. For the 1914, 1915 and 1916 cohorts, the decrement at age 90 is significantly and systematically below where we would expect it to be. This provides support for our conclusions under Graphical Diagnostic 2B that the alternative methodology for estimating exposures at age 90 and above needs revision. For the 1914 cohort, for example, the decrement at age 90 is about 1500 below where it should be. The mid-year population at age 90 is 24,842 and so the excess of 1500 equates to 6% of the cohort size at that age: again consistent with our conclusions after Graphical Diagnostic 2B.

The plots in Figure 20 show examples where the plot is not consistent with the hypothesis. For the 1919 and 1947 cohorts we can see that in the years 1992 to 2001 inclusive (grey bar), the dots are significantly and systematically below (1919) or above (1947) the reported curves of deaths. The 1919 cohort and the left-hand 1947 cohort plots are based on the inter-censal population estimates as revised after the 2011 census. Now compare the left hand 1947 plot with the right hand 1947 plot based on the post-2001-censal population estimates. Differences only exist in the years 2002 to 2010, and we can see that the inter-censal values (middle plot) have been subjected to what looks like a parallel shift down relative to the post-censal values (right-hand plot). The shift down suggests that the previous error over the 1992-2001 period has been subjected to a reversal following the 2011 census.

Similar patterns of parallel shifts exist for other cohorts including 1920, 1944, 1945 and 1946 amongst others. However, the relationship between the 1992-2001 and 2002-2010 shifts is not always similar to the pair of 1947 cohort plots.

These parallel shifts up and down are consistent with the following sequence of events:

- census to mid-year population adjustments up to 1991 seem to have produced reasonable results. The method used is not reported;
- the method for adjusting from census to mid-year outlined in Section 2.2.2 was applied in 2001;
- differences between the resulting mid-year population estimates and the post-1991-censal population estimates were then backfilled evenly over the years 1992 to 2001, producing the *signature* parallel shift down or up;

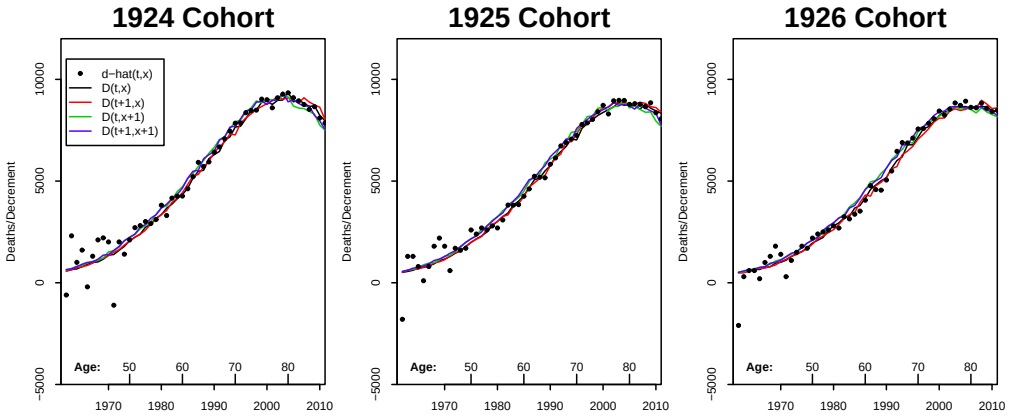

Figure 18: Death curves for the 1924 (left), 1925 (centre) and 1926 (right) cohorts by calendar year or, alternatively, by age (black and blue curves), and dots (cohort population decrements multiplied by the Covexity Adjustment Ratio). Red and green curves: death counts in adjacent cohorts adjusted for unequal cohort sizes.

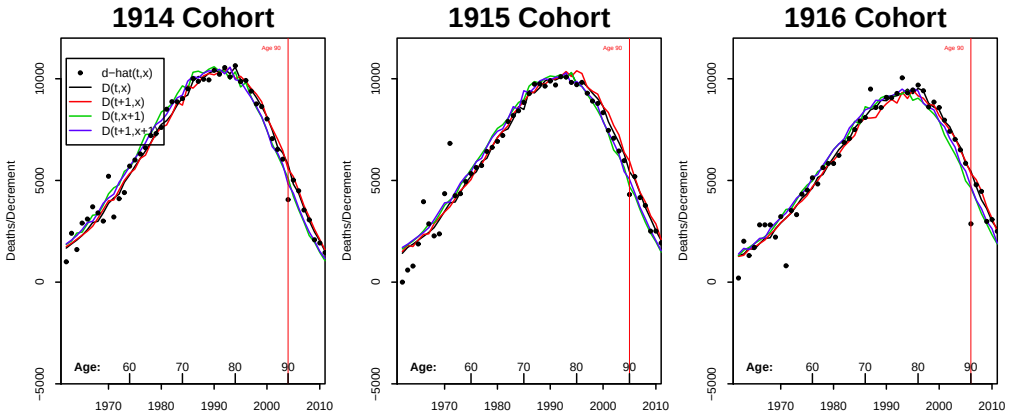

Figure 19: As Figure 18 but for the 1914 to 1916 cohorts.

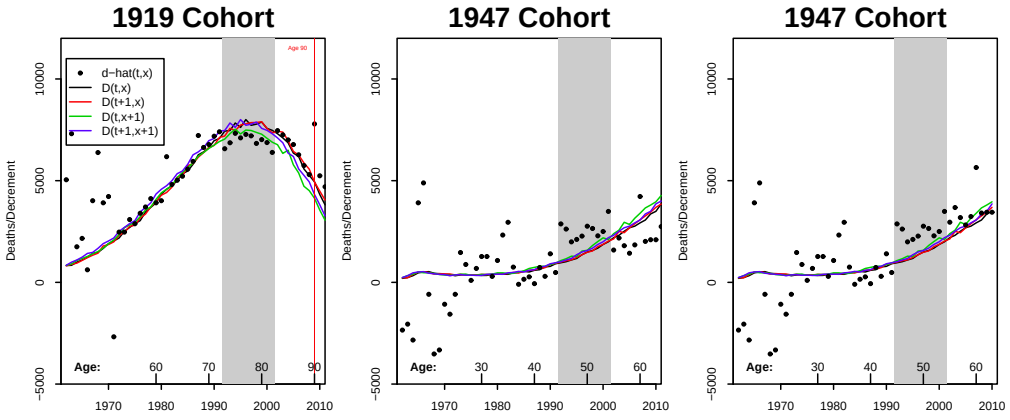

Figure 20: Death curves for the 1919 (left) and 1947 (centre and right) cohorts by calendar year or, alternatively, by age (black and blue curves), and dots (cohort population decrements multiplied by the Convexity Adjustment Ratio). Red and green curves: death counts in adjacent cohorts adjusted for unequal cohort sizes. The left and middle plots use the revised ONS data published in 2012 for years 1961 to 2011. The right plot uses the post-2001-censal ONS data for 1961 to 2010.

- the census to mid-year methodology applied in 2011 might have changed back.[19]

The method of backfilling employed by the ONS in 2001 (Duncan et al., 2002) is illustrated in a stylised way in Figure 21. This method of backfilling also produces a signature plot for Graphical Diagnostic 1 for the 1919 cohort (Figure 13): the build up of phantoms over the years 1992 to 2001 causes death rates for the 1919 cohort to drift downwards relative to neighbouring cohorts.

For the 1919 cohort, the difference between the CBD census-to-mid-year method and the method employed by the ONS amounted to 9.1% (Figure 8) or a potential error of about 6600 people. If this error did not exist in 1991 and was spread evenly over the years 1992 to 2001 then the adjustment would be 660 per year: this is entirely consistent with the magnitude of the parallel shift that we observe in Figure 20 for the 1919 cohort.

3.4 Summary

The combination of these three graphical diagnostics reveal that there remain significant anomalies in the England & Wales males population data and how it is

[19]This is difficult to establish from the data, as the older 1919 and 1920 cohorts are now above age 90 and subject to the 90+ methodology for estimating mid-year populations. Adjustments in 2011 attributable to those born in the 1940's might be due to post-censal errors in assumed net emigration.

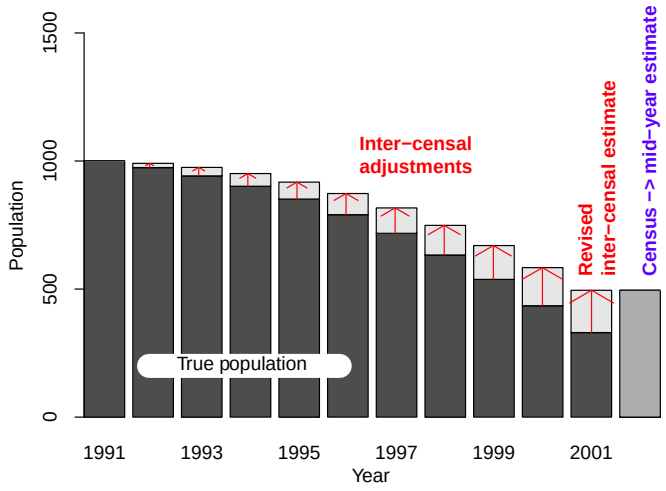

Figure 21: Stylised example of backfilling. Black bars: true population equal post-1991-censal population estimates. Dark grey bar: Mid-year population estimate following 2001 census to mid-year adjustment. Light grey bars: inter-censal adjustments build up evenly over the preceding 10 years, with zero adjustment to 1991.

used to construct crude death rates. Furthermore, although the three diagnostics are qualitatively different, they produce consistent results.

The 1919 and 1920 cohorts have been known for some time to have unusual patterns of mortality. The graphical diagnostics indicate that the anomalies are the result of unusual patterns of births and its interaction with the methodology employed in any given census year to produce mid-year population estimates.

The three diagnostics produce a stronger set of conclusions than would be possible using only one. Sometimes an anomaly will reveal itself clearly in only one of the three diagnostics and might go unobserved in the other two.

4 Model-Based Analysis of Historical Population Data

We have identified anomalies in the England & Wales males data, leading us to consider how these might be corrected in an objective way. The method we propose here uses only given sets of *exposures* and deaths data.[20] In addition, it would be possible to exploit monthly or quarterly births data as we did in our discussion of Graphical Diagnostic 2. However, we seek to develop an approach that can be applied to a wide variety of datasets, many of which will not have births data in addition to mid-year population and deaths data.

4.1 Underlying Modelling Hypotheses

In seeking to model errors in exposures, we will be guided by the following hypotheses:

- Hypothesis A: death counts are accurate;
- Hypothesis B: exposures are subject to errors, and errors following cohorts are correlated through time;
- Hypothesis C: within each calendar year the curve of underlying death rates is "smooth".

Based on these hypotheses, we adjust exposures to achieve a balance between B and C. In order to make these adjustments, we need to translate these hypotheses into a model for the errors.

4.2 Model for Errors in Exposures

We summarise here the main characteristics of the proposed model, with further details provided in Appendix D:

- for notational convenience we will assume the data runs from years 1 to n_y, covering ages 1 to n_x;
- $\hat{E}(t, x)$ = ONS estimate of the exposures for year t and age x: this is a known (published) quantity;

[20] Exposures might be equated to mid-year population estimates or, as in the case of the Human Mortality Database, data that are derived from population and deaths data. See Wilmoth et al. (2007) and Appendix B.

- $E(t,x)$ = corresponding true, but unobserveable, exposures;
- $D(t,x)$ = corresponding death count: this is assumed to be accurate;
- $m(t,x)$ = true, but unobserveable, death rate;
- define $\epsilon(t,x) = \log E(t,x)$ and $Y(t,x) = \log m(t,x)$;
- define $\phi(t,x) = \epsilon(t,x) - \hat{\epsilon}(t,x)$, where $\hat{\epsilon}(t,x)$ is the *a priori* unconditional mean of $\epsilon(t,x)$. Specifically, $\hat{\epsilon}(t,x)$ is chosen below so that the *a priori* mean of $E(t,x) = \exp[\epsilon(t,x)]$ equals $\hat{E}(t,x)$;
- conditional on $m(t,x)$ and $E(t,x)$, deaths, $D(t,x)$ have a Poisson distribution with mean and variance equal to $m(t,x)E(t,x)$;
- consistent with hypothesis B, individual time series $\phi(t+s,x+s)$ for a given (t,x) are modelled as AR(1) processes (details in Appendix D);
- consistent with hypothesis C, for a given t, $Y(t,x)$ is modelled as an ARIMA(0,3,0) process with low-variance innovations. An ARIMA(0,3,0) model with zero-variance innovations would result in a quadratic curve. With a low variance, the choice of model means that quadratic curvature is acceptable in terms of "smoothness". An ARIMA(0,2,0) process was also considered (implying that linearity was acceptable, while quadratic curvature was penalised), but we found that estimated exposure errors were not robust relative to the choice of the lowest age in the dataset (e.g., 40 or 50). By contrast, whereas results for the ARIMA(0,3,0) model were robust.

Our use of a Bayesian framework brings a number of advantages. As described above it provides a natural setting in which we can build in our prior beliefs about the error structure. The model plus data then allow us to estimate the true exposures, as well as provide a full posterior distribution for the true exposures, the true death rates, and the associated process parameters. This can then be used to make mortality forecasts that incorporate uncertainty in exposures in a very straightforward way.

The latent state variables and process parameters are estimated using Markov chain Monte Carlo (MCMC) and the Gibbs sampler (see Appendix D). This produces a vector-valued Markov chain, $M(i)$ under which, for each i, $M(i)$ records a single drawing of the latent state variables $\phi_i(t,x)$ and $Y_i(t,x)$ from the posterior distribution. This means that not only can we obtain a point estimate of each error in the exposures, but the Markov chain also gives us additional information about how much uncertainty there is around these point estimates of the errors.

Figure 22 provides a heat plot of the posterior means of the errors, $\phi(t,x)$, on the basis that the exposures are equal to the mid-year population. Greens and blues indicate that the true exposures are lower than the ONS estimates; yellows and

oranges indicate exposures need to be adjusted upwards. Mostly the errors are quite small. However, we have highlighted some (but not all) features of the plot where errors are significantly different from zero and consistently positive or negative. Most of these significant features lie on diagonals that follow specific cohorts. The 1919 and 1886 cohorts exhibit the strongest negative errors, while the strongest positive errors are for the 1900 cohort, with others associated with individual cohorts born in the 1940's, 50's and 60's. The pattern of errors for the 1919 and 1920 cohorts are entirely consistent with the conclusions that we drew from Section 3. While the earlier cohorts are either extinct or close to being so, the younger cohorts are still large in number, so any measurement errors associated with them will be materially significant.

The Bayesian analysis was repeated using exposures that had been adjusted for the uneven pattern of births using the Convexity Adjustment Ratio (CAR), as outlined in Section 2.2.4. The result of this is shown in Figure 23. The most obvious impact is that the significant negative errors for the 1920 cohort and positive errors for the 1919 cohort up to the early 1990's have now been eliminated. However, other errors, although modified, persist. For example, for the 1919 cohort in the last 15 years, errors are still significant and, from our investigation in Section 3, these errors can be attributed to a change in census-to-mid-year methodology in 2001. For the younger cohorts, the more extreme errors observable in Figure 22 have been mitigated slightly, but clearly there are additional problems with the population estimates for some of these cohorts. Adjusting for these types of errors is less easy, as at least in some cases, they reflect changes from time to time in methodology. In contrast, the Convexity Adjustment Ratio provides a straightforward, objective and constant adjustment to the exposures for each cohort. Further, a key feature of this section is that the simple model for errors and the Bayesian methodology provides us with an objective method for estimating the true exposures that was missing from Section 3.

Figure 22 also exhibits some persistent but small negative errors around ages 90 and 91, that are consistent with the concavity function in Figure 16. We can also stress that there is nothing in the model that might inadvertently produce age effects like this: what we observe is a genuine anomaly in the data. This is investigated further in Figure 24, where we plot the empirical cumulative distribution functions (ECDF) for the means of $\phi(1961, x), \ldots, \phi(2010, x)$ for ages $x = 88$ to 92. This confirms observations in Section 3 that there are some modest issues that need to be addressed in the high age methodology for estimating population and exposures above age 90.

Plots similar to Figure 24 have been considered for other age groups and some ages do stand out in a similar way. However, when we repeat the ECDF plots for two time periods 1961-1985 and 1986-2010, we see that these other age effects disappear over time (not the case for the age 88-92 plot). This points to these age-related anomalies at lower ages being the result of bias in the reporting of age at death, and reductions in this bias over time and/or improved cross checks.

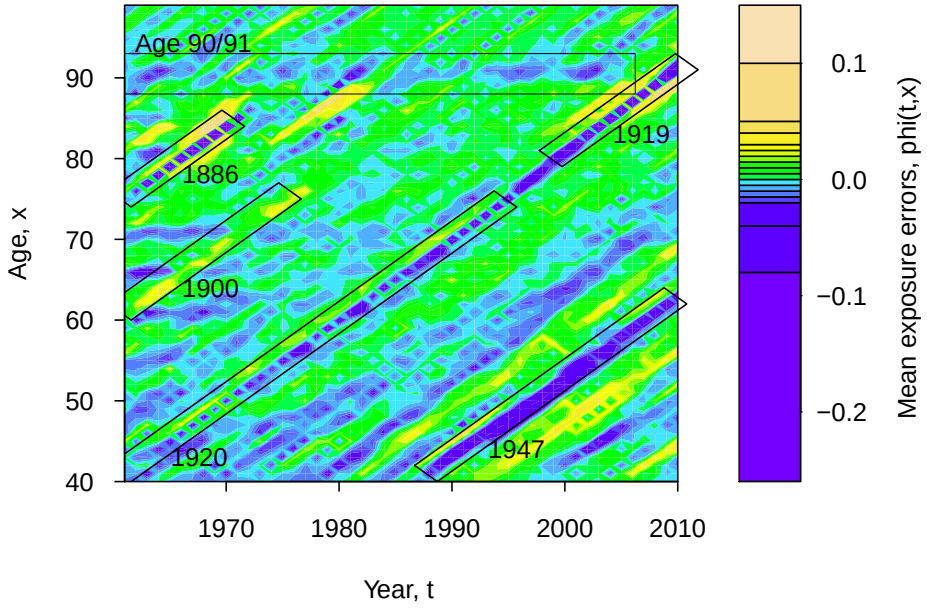

Figure 22: Heat plot of the posterior mean of the exposures error function, $\phi(t, x)$. Exposures are assumed to be equal to the mid-year population estimates.

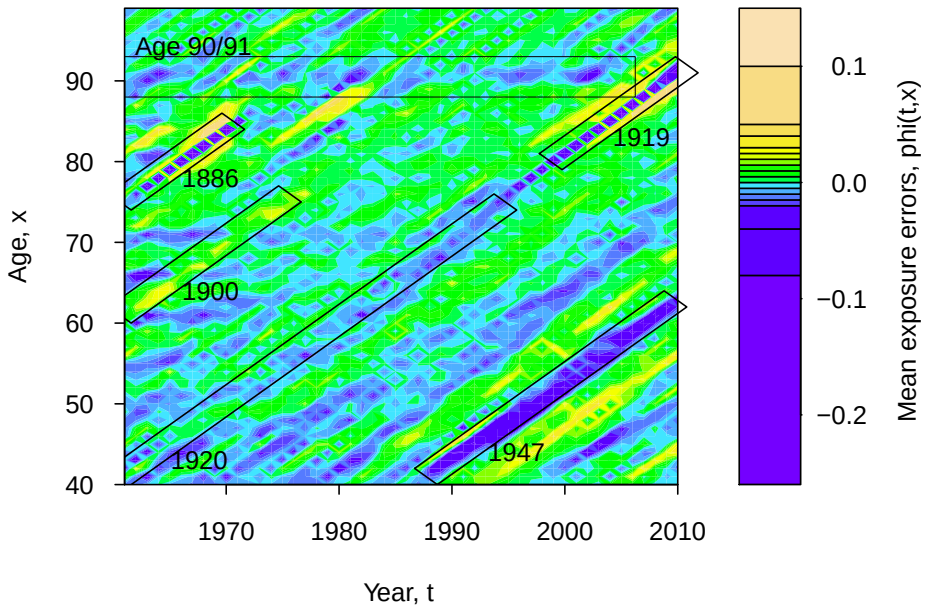

Figure 23: Heat plot of the posterior mean of the exposures error function, $\phi(t, x)$. Exposures are assumed to be equal to the mid-year population estimates multiplied by the Convexity Adjustment Ratio.

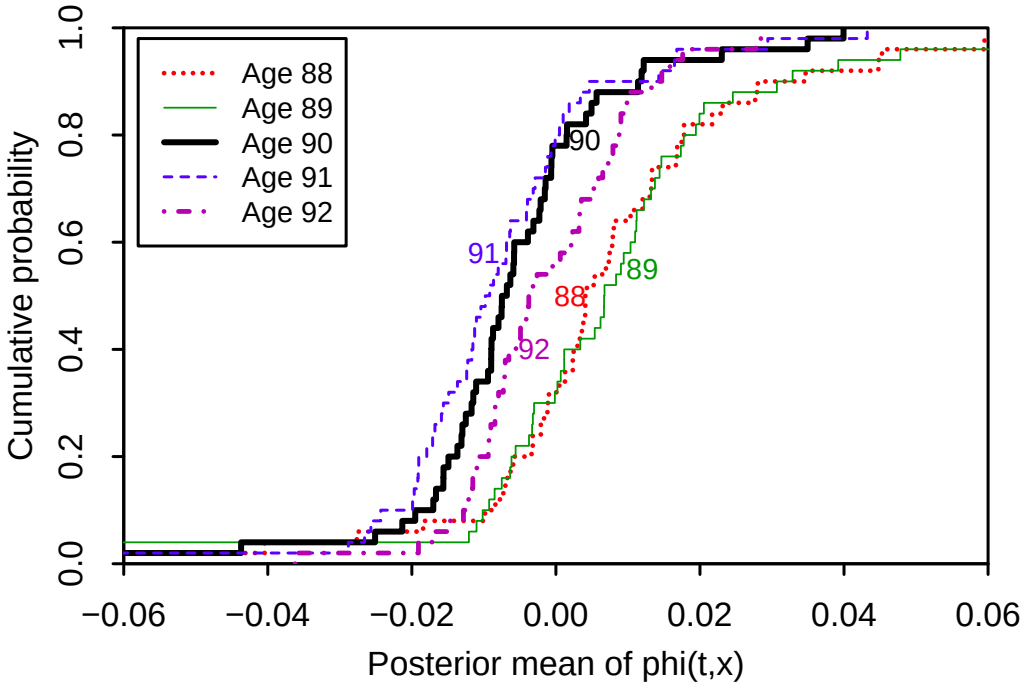

Figure 24: Empirical cumulative distributions for the posterior means of $\phi(1961, x), \ldots, \phi(2010, x)$ for ages 88 to 92.

4.3 ONS Females Data

The model was also applied to ONS England and Wales females data. The more significant estimated exposure errors in general follow a similar pattern to males (e.g., Figure 22). However, the errors for the 1919 and 1947 (and surrounding) cohorts are not quite as extreme as for males. For the 1919 cohort, a possible explanation is that, since females have lower mortality rates, the relative importance of 'phantoms that never die' grows more slowly. For the 1947 and nearby cohorts, the slightly bigger errors for males perhaps reflect greater problems in estimating the number of migrants. However, these are conjectures and other explanations might be equally valid.

5 Impact of Exposure Errors on Future Mortality Projections

We next consider what the impact is of these errors on future mortality-contingent problems. We will compare two cases:

- model-based forecasts using the unadjusted ONS exposures data (2012 revisions);
- model-based forecasts using adjusted exposures data outlined in Section 4, incorporating uncertainty in the estimated exposure errors.

Forecasts using adjusted exposures, therefore, include partial allowance for parameter uncertainty: that is, they allow for uncertainty in the exposures, but do not allow for uncertainty in estimates of, for example, model-specific period effects (see below).

By way of example, we use the following model (see Appendix C)

$$\text{logit } q(t,x) = \beta^{(0)}(x) + \kappa^{(1)}(t) + \kappa^{(2)}(t)\beta^{(2)}(x) + \kappa^{(3)}(t)\beta^{(3)}(x) + \gamma^{(4)}(t-x).$$

Mortality projections using adjusted exposures are generated using the following algorithm:

- we generate $i = 1, \ldots, N$ stochastic scenarios;
- for scenario i, draw the matrix of errors, $\phi^{(i)}(t,x)$, from the posterior distribution independent of the other scenarios;
- let $E^{(i)}(t,x) = \hat{E}(t,x)\exp[\phi^{(i)}(t,x)]$ be the matrix of adjusted exposures;
- fit the stochastic mortality model M9 to actual deaths, $D(t,x)$, and scenario i exposures, $E^{(i)}(t,x)$;
- forecast uncertainty based on scenario i exposures plus the fitted period and cohort effects under M9;
- combine the N forecasts into one set of scenarios and analyse the results.

In Figure 25, we compare mortality fan charts for unadjusted (red) and adjusted (blue) exposures data. Broadly speaking, the fans are quite similar in terms of central trajectory and spread. The similar spread indicates that parameter uncertainty in the adjusted exposures is not that large compared to uncertainty in future period and cohort effects. However, in terms of the detail the blue fans are quite smooth, while the red fans are relatively lumpy. This lumpiness is linked to the estimated

cohort effect (see Figure 26). Specifically, the blips associated with the progress through time of the 1919 and 1947 cohorts (Figure 25) correspond to jumps in the fitted cohort effect in model M9 (Figure 26). Figure 26 shows the cohort effect. The red line shows the cohort effect using the unadjusted exposures data. Mostly this is fairly smooth, but there are extreme jumps associated with the 1919 and 1947 cohorts. When the model is refitted to adjusted exposures data, the fitted cohort effect ceases to have these extreme jumps (see, for example the yellow line). Otherwise the fitted cohort effect has a similar level of smoothness as before.

We consider next the impact on cohort effects and annuity values. In Figure 27, we show, as a blue fan, the difference between the cohort effects after and before adjustment (blue fan minus the red line in Figure 26, to demonstrate the impact of the adjustment on cohort effects. Generally the differences are quite smooth and centred close to zero. However, there are some significant jumps associated with the 1919 and 1947 cohorts. We then compare these differences with the impact on annuity values (Figure 27, red line). By way of example, Figure 28 shows, by reference to cohort survivorship fan charts, why the annuity price changes by so much. Over a period of only 8 years, the proportion of survivors is 11% lower using the adjusted exposures.

From Figure 27, we can see that the impact on annuity values is strongly correlated, as we would expect, with the impact on annuity values. The expected present value of an annuity from the end of year t payable annually in arrears to a male aged x at the end of year t is

$$a(t,x) \;=\; \sum_{s=1}^{\infty}(1+r)^{-s}E[S(t,s,x)|\mathcal{M}_t] \tag{4}$$

where $S(t,s,x)$ is a survivor index that represents the proportion of males aged x at time t who survive to $t+s$, and $\mathcal{M}_t$ represents the mortality information available at time t. An interest rate of $r = 2\%$ has been assumed.

Annuity values are tabulated in Table 3 for selected cohorts, mean values are given using both unadjusted and adjusted exposures data along with the standard deviation of the random present value of the annuity (per £1 of annuity for a large population). We can see that while the 1919 cohort has the largest proportional change, the 1947 cohort has the largest nominal change, reflecting the fact that this cohort will live much longer. On the other hand, the impact of uncertain survivorship for this cohort is more distant and so its effect is dampened by the discount factor.

The 1919 cohort is the most interesting visually in much of our analyses. However, for the majority of pension plans and annuity providers, cohorts born in the 1940's and 1950's for whom we also detect significant (but smaller than 1919) anomalies should be of most concern from a financial perspective. Additionally, although price impacts are currently modest, we should take account of them, and monitor the development of the anomalies over time.

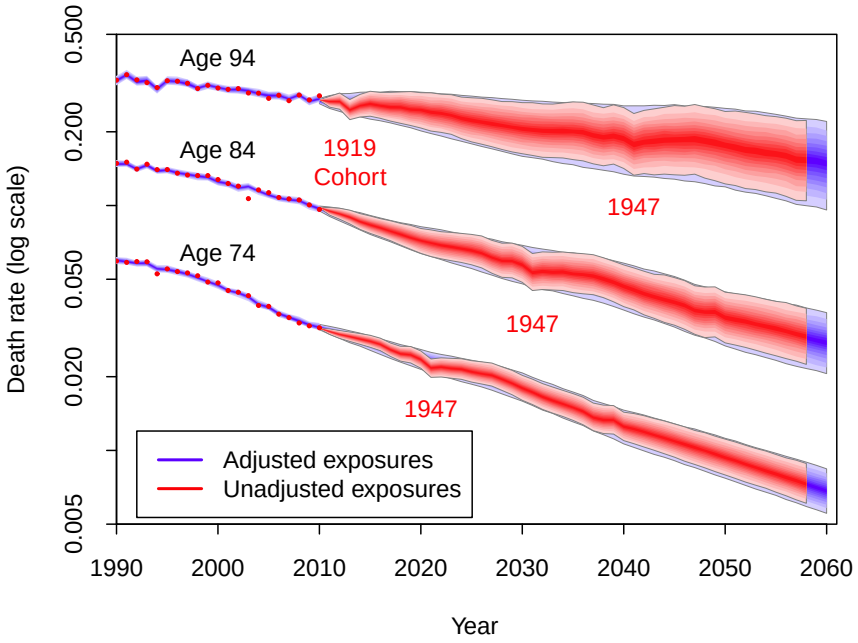

Figure 25: Mortality fan charts for ages 74, 84 and 94 for the third-generation CBD model M9. Red dots and fans: historical data and forecasts (90% prediction interval) using the unadjusted ONS exposures data. Blue fans up to 2010: adjusted exposures incorporating parameter uncertainty. Blue fans after 2010: mortality forecasts using adjusted exposures data with parameter uncertainty. Forecasts use data from 1991 to 2010.

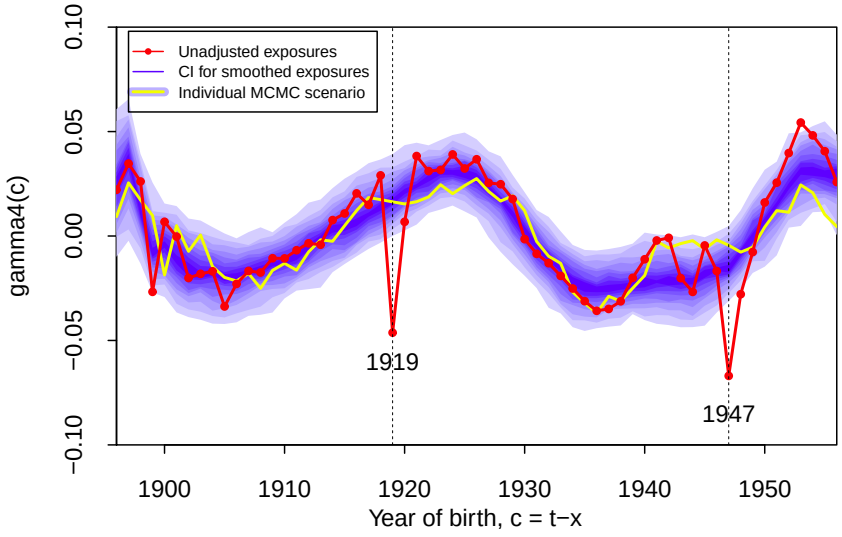

Figure 26: Cohort effect, $\gamma^{(4)}(c)$ for the third-generation CBD model M9 using unadjusted (red line) and adjusted (blue fan and yellow line) exposures data. Blue fan: 90% credibility interval (bounded by the 5% to the 95% quantiles of the posterior distribution) for $\gamma^{(4)}(c)$. Yellow line: a typical realisation of $\gamma^{(4)}(c)$ from the posterior distribution.

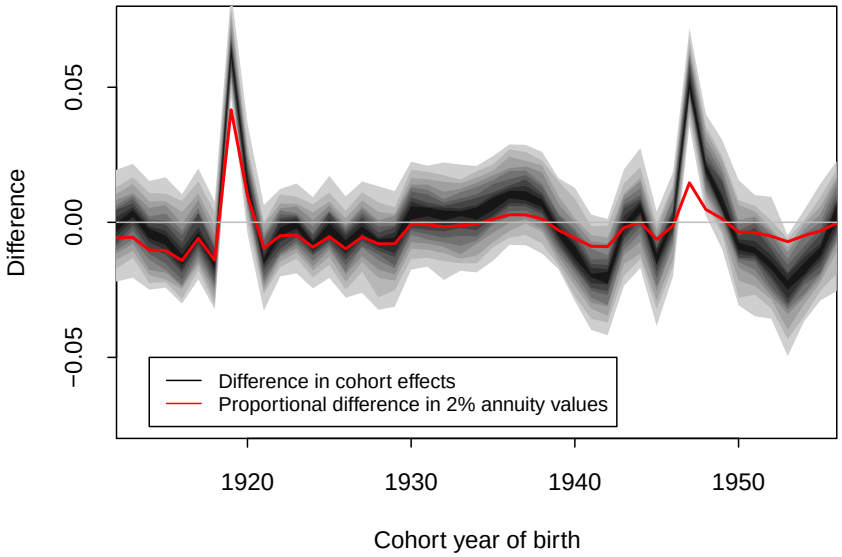

Figure 27: Comparison of changes in the estimated cohort effects versus changes in annuity prices. Grey fan: distribution of the difference between the estimated cohort effect, $\gamma_c^{(4)}$, before and after adjustment of the exposures. Red line: Impact of adjustment to exposures on annuity values for different cohorts at the end of 2010 using $r = 2\%$ as the interest rate.

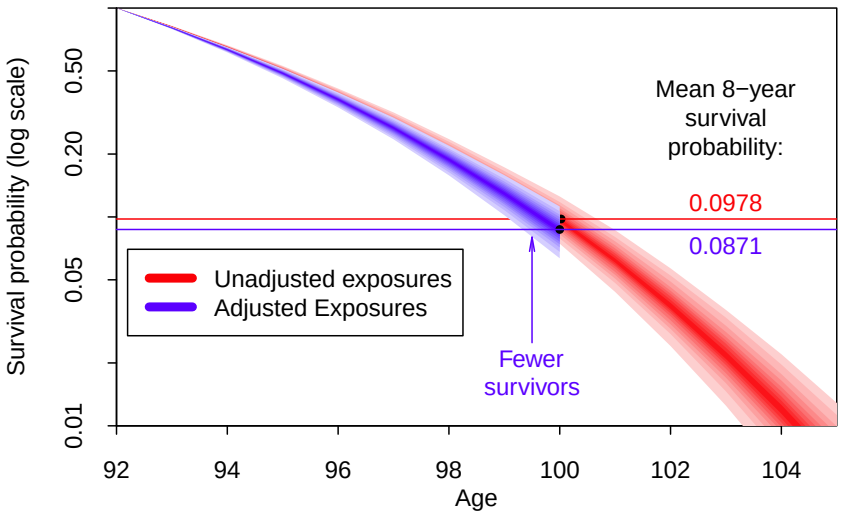

Figure 28: Survivor fan charts under the third-generation CBD model M9 for the 1919 cohort from the end of 2010 before (red fan) and after (blue) adjusting the exposures.

Age		PV: Unadjusted		PV: Adjusted	Percentage
x	Cohort	Mean	St.Dev.	Mean	Difference
63	1948	17.24	2.2%	17.16	+0.5%
64	1947	16.83	2.2%	16.58	+1.5%
65	1946	15.97	2.3%	15.99	−0.1%
92	1919	3.02	4.2%	2.90	+4.2%

Table 3: Annuity present values for various cohorts from the end of 2010, assuming $r = 2\%$, using unadjusted and adjusted exposures. Age= age at the end of 2010. Standard deviations are of $PV = \sum_{s=1}^{\infty}(1+r)^{-s}S(t,s,x)|\mathcal{M}_t$ for $t =$ end 2010 using unadjusted exposures and are expressed as a percentage of the mean. Mean present values are shown before and after adjustment of the exposures.

6 Conclusions and Next Steps

The 2011 census revisions to England & Wales population estimates by the ONS drew attention to the possibility that there are widespread errors in how population data are measured and reported in many countries. In particular, we have discovered that an uneven pattern of births within a given calendar year is a major cause of errors in population and exposures data. Different countries or agencies might derive population and exposures in different ways. But, however they do this, the estimates will be subject to potentially significant errors unless they take into account monthly or quarterly births data.

We have developed a range of methods to help identify specific errors in population, exposures and deaths data:

- graphical diagnostics providing a powerful model-free toolkit for identifying anomalies in the form of signature plots;
- a Bayesian framework enabling the size of these anomalies to be quantified;
- two-dimensional diagnostics enabling the detection of small systematic errors in exposures and deaths of less than 1%.

Our analysis using these methods has shown that errors remain in ONS population data. We have developed the Cohort-Births-Deaths (CBD) Exposures Methodology which can be used to explain many of the bigger errors. The first component was the convexity adjustment ratio which can be used to explain how persistent cohort-related errors arise when exposures are derived from mid-year population estimates. The second component was the improved method for deriving mid-year population estimates in census years from census data. Other errors were identified but only partially explained by the CBD Exposures Methodology.

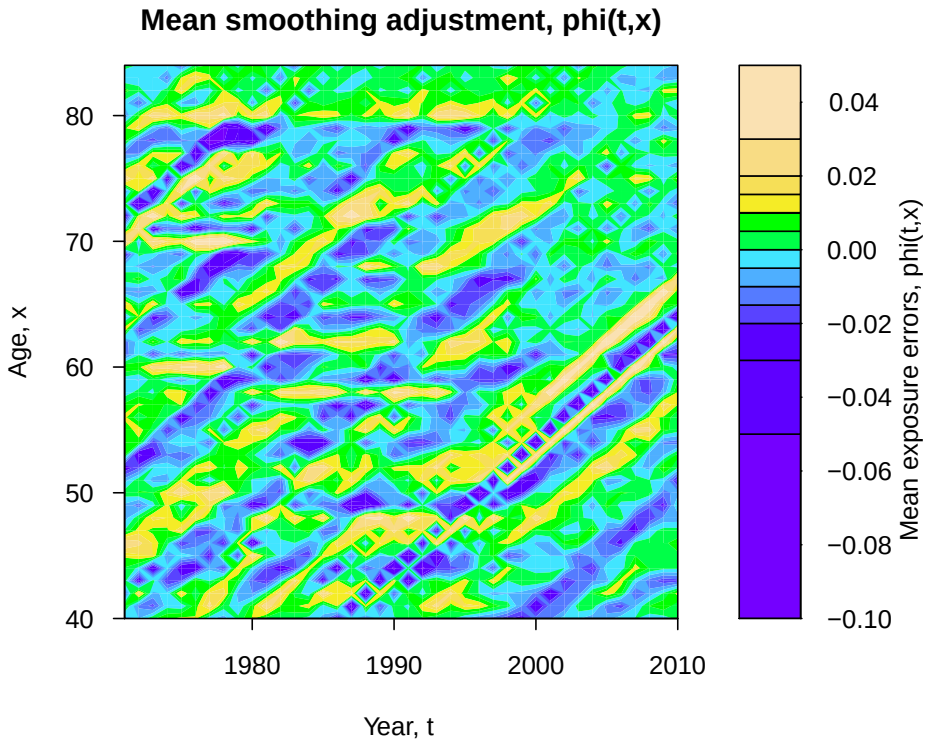

Figure 29: Heat plot of the posterior mean of the exposures error function, $\phi(t, x)$, for US males. Source data for deaths and exposures: Human Mortality Database.

In addition, anomalies in death rates have been identified that we believe are due to:

- potential small biases in the reported age at death;
- use of the Kannisto-Thatcher high age methodology resulting in a discontinuity at age 90.

Collectively, these errors can make substantial differences, particularly in respect of cohorts that are still large enough to have a significant financial impact, an example being the 1947 cohort within an annuity portfolio.

The same sources of errors – with possible variants – will apply to other countries. Some countries are similar to England & Wales in that they derive their population data from periodic (typically decadal) censuses. As one example, data collection in the US shares similar characteristics and reveals similar anomalies to England & Wales (see Figure 29). Other (typically smaller) countries have compulsory systems of registration, meaning that both deaths and population counts are accurate. However, there is still the potential for errors to be introduced when exposures are derived from annual published population values. As an example, the Netherlands

publishes accurate start of year population data. So, if average population exposures data are derived from the start of year population data, then errors can be introduced as a result of an uneven pattern of births. The implication of all this is that definitions of published data need to be read carefully and published numbers that are exact should be used wherever possible, while population figures that are derived from these (possibly by other agencies) should be used with caution.

Our analysis also suggests the following next steps:

- A fundamental review of all official mortality data and how users interpret these data.
- Until this review is complete, we need to recognise that stochastic mortality modelling is potentially flawed; in particular, we need to stop assuming exposures equal the mid-year population.
- Engagement with a range of stakeholders, including, government and supranational agencies, professional bodies, specialist longevity modelling consultancies, and private-sector holders of longevity risk (e.g., pension plans, insurers, reinsurers).
- Better communication between government agencies to exploit information contained in different public databases.
- Collection of monthly/quarterly births data (which is *essential for error mitigation*).
- Revisit the census to mid-year population methodology.
- Revisit the high-age population methodology, including comparison with that employed by the Human Mortality Database (Wilmoth et al., 2007).
- Develop software for tackling errors in population data. The software would incorporate: a module for graphical diagnostics, a model-based quantification of errors, and a (revised) high-age methodology. The output would comprise updated population and mortality tables.
- Further numerical and modelling work, such as: a review of the Kannisto-Thatcher methodology, allowance for uncertainty in reported deaths, model future revisions to data, analysis of data from other countries, and the impact on multi-population modelling.

Acknowledgements

We would like to thank the ONS for providing the deaths and population data, Stephen Richards for providing the births data and for contributing the idea that

the 1919 birth cohorts anomaly might be connected to births, and Andrew Hunt for comments on an earlier version of the paper. We would also like to thank seminar participants at the US Social Security Administration on 9 December 2013 (particularly Steve Goss, Karen Glenn, Michael Morris and Alice Wade), the Institutional Investor Breakfast Meeting, Harvard Club of New York City on 10 December 2013 (particularly Peter Nakada), and the United Nations Population Directorate on 11 December 2013 (particularly Kirill Andreev).

References

Blake, D., and Burrows, W. (2001) Survivor bonds: Helping to hedge mortality risk. *Journal of Risk and Insurance*, 68:339-348.

Blake, D., Cairns, A.J.G., and Dowd, K. (2006) Living with mortality: Longevity bonds and other mortality-linked securities. *British Actuarial Journal*, 12: 153-197.

Blake, D., Cairns, A., Coughlan, G., Dowd, D., MacMinn, R. (2013) The new life market. *Journal of Risk and Insurance*, 80: 501-558.

Booth, H., Maindonald, J., and Smith, L. (2002) Applying Lee-Carter under conditions of variable mortality decline. *Population Studies*, 56: 325-336.

Börger, M., Fleischer, D., and Kuksin. N. (2014) Modeling mortality trends under modern solvency regimes. *ASTIN Bulletin*, 44:1-38.

Brouhns, N., Denuit, M., and Vermunt J.K. (2002) A Poisson log-bilinear regression approach to the construction of projected life tables. *Insurance: Mathematics and Economics*, 31:373-393.

Cairns, A.J.G. (2011) Modelling and management of longevity risk: Approximations to survival functions and dynamic hedging. *Insurance: Mathematics and Economics*, 49: 438-453.

Cairns, A.J.G. (2013) Robust hedging of longevity risk. *Journal of Risk and Insurance*, 80: 621-648.

Cairns, A.J.G. (2014) Modelling and management of longevity risk. Chapter TBA in *TBA*, Olivia S. Mitchell, editor.

Cairns, A.J.G., Blake, D., and Dowd, K. (2006) A two-factor model for stochastic mortality with parameter uncertainty: Theory and calibration. *Journal of Risk and Insurance,* 73: 687-718.

Cairns, A.J.G., Blake, D., and Dowd, K. (2008) Modelling and management of mortality risk: A review. *Scandinavian Actuarial Journal*, 2008: 79-113.

Cairns, A.J.G., Blake, D., Dowd, K., Coughlan, G.D., Epstein, D., Ong, A., and Balevich, I. (2009) A quantitative comparison of stochastic mortality models using

data from England & Wales and the United States. *North American Actuarial Journal*, 13: 1-35.

Cairns, A.J.G., Blake, D., Dowd, K., Coughlan, G.D., Epstein, D., and Khalaf-Allah, M. (2011a) Mortality density forecasts: An analysis of six stochastic mortality models. *Insurance: Mathematics and Economics*, 48: 355-367.

Cairns, A.J.G., Blake, D., Dowd, K., Coughlan, G.D., and Khalaf-Allah, M. (2011b) Bayesian stochastic mortality modelling for two populations. *ASTIN Bulletin*, 41: 29-59.

Cairns, A.J.G., Blake, D., Dowd, K., and Coughlan, G.D. (2013) Longevity hedge effectiveness: A decomposition. *Quantitative Finance*, 14: 217-235.

Coughlan, G., Epstein, D., Sinha, A., and Honig, P. (2007) *q*-forwards: Derivatives for transferring longevity and mortality risk. JP Morgan, London.

Currie, I.D. (2011) Modelling and forecasting the mortality of the very old. *ASTIN Bulletin*, 41: 419-427.

Czado, C., Delwarde, A., and Denuit, M. (2005) Bayesian Poisson log-bilinear mortality projections. *Insurance: Mathematics and Economics*, 36: 260-284.

Dahl, M., Melchior, M., and Møller, T. (2008) On systematic mortality risk and risk minimisation with survivor swaps. *Scandinavian Actuarial Journal*, 2008(2-3): 114-146.

Dowd, K., Cairns, A.J.G., Blake, D., Coughlan, G.D., and Khalaf-Allah, M. (2011) A gravity model of mortality rates for two related populations. *North American Actuarial Journal*, 15: 334-356.

Duncan, C., Chappell, R., Smith, J., Clark, L., and Ambrose, F. (2002) Rebasing the annual midyear population estimates for England and Wales. *Population Trends* 109. Office for National Statistics.

Hunt, A., and Blake, D. (2014) A general procedure for constructing mortality models. *North American Actuarial Journal*, 18: 116-138.

Hyndman, R.J., and Ullah, M.S. (2007) Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics and Data Analysis*, 51: 4942-4956.

Jarner, S.F., and Kryger, E.M. (2011) Modelling adult mortality in small populations: The SAINT model. *ASTIN Bulletin*, 41: 377-418.

Lee, R.D., and Carter, L.R. (1992) Modeling and forecasting U.S. mortality. *Journal of the American Statistical Association*, 87: 659-675.

Li, J.S.-H., Hardy, M.R., and Tan, K.S. (2009) Uncertainty in model forecasting: An extension to the classic Lee-Carter approach. *ASTIN Bulletin,* 39: 137-164.

Li, J.S.-H., and Luo, A. (2012) Key q-duration: A framework for hedging longevity

risk. *ASTIN Bulletin*, 42: 413-452.

Li, J.S.-H., and Hardy, M.R. (2011) Measuring basis risk in longevity hedges. *North American Actuarial Journal*, 15: 177-200.

Li, N., and Lee, R. (2005) Coherent mortality forecasts for a group of populations: An extension of the Lee-Carter method. *Demography*, 42(3): 575-594.

Mavros, G., Cairns, A.J.G., Streftaris, G., and Kleinow, T. (2014) A parsimonious approach to stochastic mortality modelling with dependent residuals. Working paper, Heriot-Watt University.

ONS (2002a) *Mortality Statistics – General – Review of the Registrar General on deaths in England & Wales, 2000, Series DH1 No.33.* HMSO, London.

ONS (2002b) *Census 2001: First results on population for England and Wales.* TSO, London.

ONS (2002c) Methods used to revise the 1982-2000 annual mid-year population estimates for England and Wales. Office for National Statistics.

ONS (2012a) Explaining the Difference between the 2011 Census Estimates and the Rolled-Forward Population Estimates. Office for National Statistics, 16 July 2012.

ONS (2012b) Population Estimates for England & Wales , Mid-2002 to Mid-2010 Revised (National). Office for National Statistics, 13 December 2012.

ONS (2013) Quality and methodology information. Office for National Statistics, 10 July 2013.

ONS (2014) Calculating population estimates of the very elderly. Office for National Statistics. www.ons.gov.uk. Accessed 9/5/2014.
http://www.ons.gov.uk/ons/guide-method/method-quality/specific/population-and-migration/pop-ests/calculating-estimates-of-the-very-elderly/index.html

Plat, R. (2009) On stochastic mortality modeling. *Insurance: Mathematics and Economics*, 45, 393404.

Renshaw, A.E., and Haberman, S. (2006) A cohort-based extension to the Lee-Carter model for mortality reduction factors. *Insurance: Mathematics and Economics,* 38: 556-570.

Richards, S.J. (2008) Detecting year-if-birth mortality patterns with limited data. *Journal of the Royal Statistical Society, Series A*, 171: 279-298.

Thatcher, R., Kannisto, V., and Andreev, K. (2002) The survivor ratio method for estimating numbers at high ages. *Demographic Research* 6: 1-18.

Willets, R.C. (2004) The cohort effect: Insights and explanations. *British Actuarial Journal*, 10: 833-877.

Wilmoth, J.R., Andreev, K., Jdanov, D., and Glei, D.A. (2007) Methods protocol

for the Human Mortality Database. See www.mortality.org. (Accessed 16/3/2014.)

A Estimating Age 0 Population and Exposures from Births Data

UK Office for National Statistics (ONS) data for numbers of male and female births has been sourced by quarter of birth from 1838 to 1937, and by month from 1938 to 2003 inclusive (166 years).

Throughout this appendix we will assume that after birth there are no deaths at age 0 and no migration.

A.1 Quarterly births data

With quarterly births data (1838 to 1937), we have $B_Q(t)$ = births in the quarter t to $t+0.25$, where t is itself an integer multiple of 0.25. For the period 1938 to 2003, we simply aggregate the monthly births data into quarterly data.

At quarterly intervals, the population aged 0 last birthday at time t is then known exactly and is the sum of the previous four quarterly birth counts. Thus, if t is a multiple of 0.25,

$$P(t,0) = B_Q(t-1) + B_Q(t-0.75) + B_Q(t-0.5) + B_Q(t-0.25).$$

For times t in between quarter dates, we suggest linear interpolation, although other methods involving local fitting of polynomials or splines could be considered. Linear interpolation is equivalent to the assumption that quarterly births are uniformly distributed over the three months.

A.2 Monthly births data

With monthly births data (1938 to 2003), we have $B_M(t)$ = births in the month t to $t+1/12$ where t is itself an integer multiple of 1/12. For the period 1838 to 1937, we make the simple assumption that quarterly births are spread uniformly across each of the three months.

At monthly intervals, the population aged 0 last birthday at time t is then known exactly and is the sum of the previous twelve monthly birth counts. Thus, if t is a multiple of 1/12,

$$P(t,0) = \sum_{i=1}^{12} B_M(t-i/12).$$

For times t in between each month end, we suggest linear interpolation, although other methods involving local fitting of polynomials or splines could be considered.

The use of monthly data is clearly better than quarterly where these data are available. Furthermore, where source data are quarterly, the proposed linear interpolation makes age 0 population estimates under the monthly and quarterly give identical results. Consequently, we will use monthly births data to calculate population estimates.

A.3 Exposures

Lastly, we seek to estimate the exposures for age 0 last birthday in year t. We have $E(t,0) = \int_0^1 P(t+s,0)ds$ where the function $P(t+s,0)$ has been estimated using the monthly births data as above. We then use Simpson's Rule for numerical integration:

$$E(t,0) \;=\; \sum_{i=0}^{12} w_i P(t+i/12,0)$$

where $w_i = 1/36$ for $i = 0, 12$, $w_i = 4/36$ for $i = 1, 3, 5, 7, 9, 11$ and $w_i = 2/36$ for $i = 2, 4, 6, 8, 10$. Equivalently,

$$E(t,0) \;=\; \sum_{i=-12}^{11} b_i B_M(t+i/12)$$

where $b = (1, 5, 7, 11, 13, 17, 19, 23, 25, 29, 31, 35, 35, 31, 29, 25, 23, 19, 17, 13, 11, 7, 5, 1)/36$.[21]

B Human Mortality Database Estimation of Exposures

We summarise here the main elements of Wilmoth et al. (2007) that concern the calculation of exposures over the majority of ages (i.e., not age 0, and not very high ages):

- If 1 January population estimates are not published then these are derived using linear interpolation. For example, in England & Wales (EW), mid-year population estimates, $P(t+\frac{1}{2},x)$, are published. Linear interpolation means that $P(t,x) = \frac{1}{2}\left(P(t-\frac{1}{2},x) + P(t+\frac{1}{2},x)\right)$.

[21] Note the elements of b sum to 12 rather than 1 as the births are monthly rather than annual.

- Exposures are then (for EW)

$$\begin{aligned} E(t,x) &= \frac{1}{2}\left(P(t,x)+P(t+1,x)\right)+\frac{1}{6}\left(D_L(t,x)-D_U(t,x)\right) \\ &= \frac{1}{4}P(t-\frac{1}{2},x)+\frac{1}{2}P(t+\frac{1}{2},x)+\frac{1}{4}P(t+\frac{3}{2},x)+\frac{1}{6}\left(D_L(t,x)-D_U(t,x)\right) \end{aligned}$$

where $D_L(t,x)$ and $D_U(t,x)$ are the lower and upper Lexis triangles of deaths. $D_L(t,x)$ = deaths in year t aged x last birthday, and born in year $t-x$, $D_U(t,x)$ = deaths in year t aged x last birthday, and born in year $t-x-1$, and $D_L(t,x)+D_U(t,x) = D(t,x)$. Typically, $D_L(t,x)$ is a bit less than $D_U(t,x)$ (the $t-x-1$ birth cohort are older and, therefore, experience slightly higher mortality rates), but the method proposed by Wilmoth et al. (2007, equations (4) and (7)) adjusts the balance further to take account of age and, more importantly, relative birth counts in years $t-x-1$ and $t-x$. Birth counts are by calendar year rather than by month or quarter.

C The M9 Stochastic Mortality Model

By way of example in this paper, we have used the following model labelled M9 (an extension of model M7 proposed by Cairns et al., 2009, and similar in motivation to that of Plat, 2009)

$$\text{logit } q(t,x) = \beta^{(0)}(x) + \kappa^{(1)}(t) + \kappa^{(2)}(t)\beta^{(2)}(x) + \kappa^{(3)}(t)\beta^{(3)}(x) + \gamma^{(4)}(t-x)$$

where the $\kappa^{(i)}(t)$ are stochastic period effects, $\gamma^{(4)}(c)$ is a stochastic cohort effect, $\beta^{(2)}(x) = (x-\bar{x})$ and $\beta^{(3)}(x) = (x-\bar{x})^2 - \sigma_x^2$ are fixed age effects, $\bar{x}$ and σ_x are the mean and standard deviation respectively of the ages under consideration, and $\beta^{(0)}(x)$ is an additional non-parametric age effect (borrowing from the Lee-Carter approach). As with Plat (2009), the introduction of $\beta^{(0)}(x)$ allows the model to be applied with greater accuracy to a wider range of ages than would normally be recommended for the first and second-generation CBD mortality models (M5, M6 and M7).

M9 has seven identifiability constraints: $\sum_t \kappa^{(i)}(t) = 0$ for $i = 1, 2, 3$ and $\sum_c \gamma^{(4)}(c)c^k = 0$ for $k = 0, 1, 2, 3$.

D Modelling Uncertainty in Exposures

We assume that, for each t and x, $E(t,x)/\hat{E}(t,x)$, the ratio of true to published exposures, has a log-normal distribution with mean 1: that is, there is no *a priori* reason to assume that the ONS would deliberately over or under-estimate exposures.

Additionally, we model the logarithm of the latent death rates, $Y(t,x) = \log m(t,x)$, as having an underlying smoothness across ages in a given calendar year, t. Basic notation is defined in Section 4.

The components of the modelling process are as follows:

- Define $\epsilon(t,x) = \log E(t,x)$ and $Y(t,x) = \log m(t,x)$.
- Define $\phi(t,x) = \epsilon(t,x) - \hat{\epsilon}(t,x)$ where $\hat{\epsilon}(t,x)$ is the prior *unconditional* mean of $\epsilon(t,x)$. $\hat{\epsilon}(t,x)$ is chosen below so that $E(t,x) = \exp[\epsilon(t,x)]$ has prior unconditional mean $\hat{E}(t,x)$.
- Conditional on $m(t,x)$ and $E(t,x)$, deaths, $D(t,x)$, have a Poisson distribution with mean and variance equal to $m(t,x)E(t,x)$.
- Model for the propagation of errors in exposures.
  - Let $0 < \theta < 1$, and $\sigma_\phi > 0$ be the model parameters.
  - Let $Z_\phi(t,x)$ for $t = 2,\ldots,n_y$ and $x = 1,\ldots,n_x$ be independent and identically distributed standard normal random variables.
  - For $t = 2,\ldots,n_y$ and $x = 2,\ldots,n_x$:

    $$\epsilon(t,x) \;=\; \hat{\epsilon}(t,x) \;+\; \theta\,(\epsilon(t-1,x-1) - \hat{\epsilon}(t-1,x-1)) + \sigma_\phi Z_\phi(t,x)$$

    or, more simply,

    $$\phi(t,x) \;=\; \theta\phi(t-1,x-1) + \sigma_\phi Z_\phi(t,x).$$

  - For $t = 2,\ldots,n_y$ and $x = 1$:

    $$\phi(t,1) = \theta\phi(t-1,1) + \sigma_\phi Z_\phi(t,1).$$

  - Thus, if we write $\phi(t) = (\phi(t,1),\ldots,\phi(t,n_x))'$, then, for $t = 2,\ldots,n_y$, $\phi(t)$ has a multivariate normal distribution with mean $A\phi(t-1)$ and covariance matrix V_ϕ, where

$$A = \begin{pmatrix} \theta & 0 & 0 & \ldots & & 0 \\ \theta & 0 & \ldots & & & \\ 0 & \theta & 0 & \ldots & & \\ 0 & 0 & \theta & 0 & \ldots & \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & \ldots & 0 & \theta & 0 \end{pmatrix}, \quad \text{and} \quad V_\phi = \begin{pmatrix} \sigma_\phi^2 & 0 & \ldots & 0 \\ 0 & \sigma_\phi^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & \sigma_\phi^2 \end{pmatrix}$$

– Let $\phi(1) = (\phi(1,1), \ldots, \phi(1,n_x))'$ be the vector of errors in year 1. $\phi(1)$ has a multivariate normal distrubution with mean vector 0 and covariance matrix $V_{0\phi}$, this being the stationary distribution of the vector autoregressive (VAR) process for $\phi(t)$, and is independent of the $Z_\phi(t,x)$ for $t > 1$.

For the simple VAR process we have:

$$Var(\phi(1,x)) = \sigma_\phi^2/(1-\theta^2) \quad \text{for all } x$$

and, for $y > x$,

$$\text{cor}\,(\phi(1,x), \phi(1,y)) = \theta^{2(y-1)}$$

which, in combination, define the covariance matrix $V_{0\phi}$. The reality concerning the introduction and propagation of errors in exposure estimates is, no doubt, much more complex than this VAR model. Our choice has been kept deliberately simple in the hope that the data themselves will prevail and that the VAR model will not influence strongly the actual estimates of the true exposures.

- Model for the underlying death rates.

  Individual years are treated in isolation and assume the following:

$$Y(t,x) \;=\; 3Y(t,x-1) - 3Y(t,x-2) + Y(t,x-3) + \sigma_Y Z_Y(t,x)$$

  where the $Z_Y(t,x)$ are i.i.d. standard normal random variables. Thus, for a given t, $Y(t,x)$ is an ARIMA(0,3,0) process, giving locally quadratic predictions plus noise proportional to σ_Y.

  Specifically, we choose σ_Y to be small so that, for each t, the observed process $Y(t,x)$ prefers to be smooth and linear rather than rough or curved. The choice of an ARIMA(0,3,0) was preferred to an ARIMA(0,2,0) (linear prediction plus noise) because the former resulted in a much greater degree of robustness in the exposures errors relative to the choice of minimum age.

- Model for deaths.

  For each t and x, and given $\phi(t,x)$ and $Y(t,x)$, we assume that deaths, $D(t,x)$, have a log-normal distribution. $d(t,x) = \log D(t,x)$ thus has a normal distribution and we define its mean to be $\hat{\epsilon}(t,x) + \phi(t,x) + Y(t)$ (which is the log of the mean of $D(t,x)$ under the alternative conditional Poisson assumption) and its variance to be $S_D(t,x) = 1/\hat{D}(t,x)$, where $\hat{D}(t,x)$ is set equal to the observed deaths. The rationale for this choice is that

  – $E[\exp d(t,x)] \approx \exp[\hat{\epsilon}(t,x)+\phi(t,x)+Y(t)]$ and $Var[\exp d(t,x)] \approx \exp[\hat{\epsilon}(t,x)+\phi(t,x)+Y(t)]$: this is consistent with the conditional Poisson assumption.

  - Fixing $S_D(t,x)$ in the self referential way that we have is consistent with the idea that if we were to resimulate a Poisson random variable based on a single observation, we might simulate a Poisson with mean (and variance) equal to the value of that one observation. This is a good approximation if numbers of deaths are relatively large.

- Apart from the time-series structures specified above, the matrices of latent state variables, Y and ϕ, are assumed to have uninformative, uniform prior distributions on $(-\infty,\infty)^{n_Y \times n_x}$.

- We treat the process parameters as exogenously specified: $\sigma_Y = 0.01$, $\sigma_V = 0.02$ and $\theta = 0.9$. $\sigma_Y = 0.01$ reflects our desire for a high degree of smoothness in $Y(t,x)$. $\sigma_V = 0.02$ reflects the revisions that we have seen in the ONS exposures data. $\phi = 0.9$ reflects the strong persistence of errors in the exposures data. Estimated exposure errors have been tested for sensitivity to these three parameters and results have been found to be robust.

The log-likelihood for our model is then:

$$f(\phi, Y|d,\hat{E}) = f_1(d|Y,\phi,\hat{E}) + f_2(\phi) + f_3(Y) \tag{5}$$

where

$$d(t,x) = \log D(t,x)$$

$$f_1(d|Y,\phi,\hat{E}) = -\frac{1}{2}\sum_{t=1}^{n_y} (d(t) - Y(t) - \hat{\epsilon}(t) - \phi(t))' S_D(t)^{-1} (d(t) - Y(t) - \hat{\epsilon}(t) - \phi(t)) + \text{const.}$$

$$f_2(\phi) = f_{2A}(\phi(1)) + \sum_{t=2}^{n_1} f_{2B}(\phi(t))$$

$$f_{2A}(\phi(t)) = -\frac{1}{2}(\phi(t) - A\phi(t-1))' V_\epsilon^{-1} (\phi(t) - A\phi(t-1)) + \text{const.}$$

$$f_{2B}(\phi(1)) = -\frac{1}{2}\phi(1)' V_{0\epsilon}^{-1} \phi(1) + \text{const.}$$

and

$$f_3(Y) = -\frac{1}{2}\sum_{t=1}^{n_y} \frac{1}{\sigma_Y^2} Y(t)' \Delta' \Delta Y(t) + \text{const.}$$

and where

$$\Delta = \begin{pmatrix} 1 & -3 & 3 & -1 & 0 & \cdots & \cdots \\ 0 & 1 & -3 & 3 & -1 & 0 & \cdots \\ \vdots & 0 & 1 & -3 & 3 & -1 & \cdots \\ \vdots & & \ddots & \ddots & \ddots & \ddots & \ddots \\ & & 0 & 1 & -3 & 3 & -1 \end{pmatrix} \quad \text{(an } n_x - 3 \times n_x \text{ matrix).}$$

Since Y and ϕ have uniform prior distributions, the log-likelihood above (Equation 5) is equal to the log-posterior density. The posterior distribution for the combined

(Y, ϕ) can be seen to be multivariate normal. However, the dimension of the state space $(2n_x n_y)$ is very large, and so identifying the mean and covariance matrix for the posterior is computationally almost impossible. Instead, we adopt a Markov chain Monte Carlo (MCMC) approach using the Gibbs sampler as a proposal distribution. The MCMC algorithm proceeds as follows:

- Let $Y(t)^C$ represent all elements of (Y, ϕ) apart from the vector $Y(t)$, and $\phi(t)^C$ represent all elements of (Y, ϕ), apart from the vector $\phi(t)$.
- For $t = 1, \ldots, n_y$, the conditional posterior for $Y(t)$ given $Y(t)^C$ is multivariate normal with mean $\mu_Y \equiv \mu(Y(t)^C)$ and covariance matrix $H_Y \equiv H(Y(t)^C))$ where

$$\begin{aligned} H_Y &= S_D(t)^{-1} + \frac{1}{\sigma_Y^2}\Delta_2'\Delta_2 \\ \mu_Y &= H_Y^{-1} S_D(t)^{-1} \left(d(t) - \hat{\epsilon}(t) - \phi(t)\right). \end{aligned}$$

- The conditional posterior distribution for $\phi(1)$ given $\phi(1)^C$ is multivariate normal with mean $\mu_\phi(1, \phi(1)^C)$ and covariance matrix $H_\phi(1, \phi(1)^C)$ where

$$\begin{aligned} H_\phi &= S_D(1)^{-1} + A' V_\epsilon^{-1} A + V_{0\phi}^{-1} \\ \mu_\phi &= H_\phi^{-1} \left(S_D(1)^{-1}(d(1) - \hat{\epsilon}(1) - Y(1)) + A' V_\phi^{-1} \phi(2)\right). \end{aligned}$$

- The conditional posterior distribution for $\phi(t)$ given $\phi(t)^C$, for $t = 2, \ldots, n_y - 1$, is multivariate normal with mean $\mu_\phi(t, \phi(t)^C)$ and covariance matrix $H_\phi(t, \phi(t)^C)$ where

$$\begin{aligned} H_\phi &= S_D(t)^{-1} + V_\phi^{-1} + A' V_\phi^{-1} A \\ \mu_\phi &= H_\phi^{-1} \left(S_D(t)^{-1}(d(t) - \hat{\epsilon}(t) - Y(t)) + V_\phi^{-1} A \phi(t-1) + A' V_\phi^{-1} \phi(t+1)\right). \end{aligned}$$

- The conditional posterior distribution for $\phi(n_y)$ given $\phi(n_y)^C$ is multivariate normal with mean $\mu_\phi(n_y, \phi(n_y)^C)$ and covariance matrix $H_\phi(n_y, \phi(n_y)^C)$ where

$$\begin{aligned} H_\phi &= S_D(n_y)^{-1} + V_\phi^{-1} \\ \mu_\phi &= H_\phi^{-1} \left(S_D(n_y)^{-1}(d(n_y) - \hat{\epsilon}(n_y) - Y(n_y)) + V_\phi^{-1} A \phi(n_y - 1)\right). \end{aligned}$$